ESTIMATING FEATURES OF A DISTRIBUTION FROM BINOMIAL DATA*

Arthur Lewbel+
Boston College

Oliver Linton†
London School of Economics

Daniel McFadden‡
University of California, Berkeley

Discussion paper
No. EM/2006/507
September 2006

The Suntory Centre
Suntory and Toyota International Centres for Economics and Related Disciplines
London School of Economics and Political Science
Houghton Street
London WC2A 2AE
Tel: 020 7955 6679

---

* This research was supported in part by the National Science Foundation through grants SES-9905010 and SBR-9730282, by the E. Morris Cox Endowment, and by the ESRC.
+ Department of Economics, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467, USA. Phone: (617) 552-3678. E-mail address: lewbel@bc.edu
† Department of Economics, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom. E-mail address: lintono@lse.ac.uk
‡ Department of Economics, University of California, Berkeley, CA 94720-3880, USA. E-mail address: mcfad-den@econ.berkeley.edu

Abstract

A statistical problem that arises in several fields is that of estimating the features of an unknown distribution, which may be conditioned on covariates, using a sample of binomial observations on whether draws from this distribution exceed threshold levels set by experimental design. Applications include bioassay and destructive duration analysis. The empirical application we consider is referendum contingent valuation in resource economics, where one is interested in features of the distribution of values (willingness to pay) placed by consumers on a public good such as endangered species. Sample consumers are asked whether they favor a referendum that would provide the good at a cost specified by experimental design. This paper provides estimators for moments and quantiles of the unknown distribution in this problem under both nonparametric and semiparametric specifications.

JEL Codes: C14, C25, C42, H41.

Keywords: Willingness to Pay, Contingent Valuation, Discrete Choice, Bi-nomial response, Bioassay, Destructive Duration Testing, Semiparametric, Nonparametric, Latent Variable Models.

1 Introduction

We consider estimation of moments and quantiles of an unknown distribution, which may be conditioned on covariates, using a sample of binomial observations on whether draws from this distribution exceed threshold levels set by experimental design. For example, consider contingent valuation studies, which are used to assess the willingness to pay (WTP) for a good or resource, such as a change in environmental quality. In referendum format experiments each subject is asked if their WTP exceeds a test value or bid v chosen by experimental design. Objects of interest to estimate from subject's binary responses include the mean, variance and (for median voter models) median WTP across individuals, perhaps conditioned on characteristics x that make them likely voters.[1]

The problem can be generally described as uncovering features of conditional survival curves from unbounded interval censored data. Let W denote a random failure time, and let $G(w \mid x) = \Pr(W > w \mid X = x)$ denote the survival curve conditioned on a d vector of time invariant covariates x. In duration experiments w is time, but it could also be, e.g., an administered dose of a toxin where W is the lethal dose and G is the dose response curve, or W could be WTP as above. Assume we have interval censored data, that is, for each individual i, W_i is latent and we observe (X_i, V_i, Y_i) where V_i is a test level in the w dimension set by experimental design, V_i is drawn independent of W_i given a covariate vector X_i, and $Y_i = 1(W_i > V_i)$ is a binary indicator for the event $W_i > V_i$. The important attributes of these data are the experimental design feature that W and V are conditionally independent given X, and that failure times W are not observed, so properties of the distribution G, such as the mean and other failure time moments, must be inferred from the binary status indicators Y_i and single test levels V_i. Our estimators could be readily extended to multiple (adaptive) test levels and multinomial status.

In the contingent valuation example, y equals one if the subject's WTP W exceeds the bid v chosen by experimental design. More generally, y could indicate if a benefit W exceeds a posited cost v in, e.g., an experimental voting or investment decision. In bioassay, v is the time an animal

[1] Other experimental designs include follow up queries to gain more information about WTP, and open ended questions, where subjects are simply asked to state their WTP. Open ended questions often suffer from high rates of nonresponse (with possible selection bias), while referendum format follow up responses can be biased due to the framing effect of the first bid. This shadowing effect is common in unfolding bracket survey questions. See McFadden (1994) for references and experimental evidence regarding response biases. Other issues regarding the framing of questions also impact survey responses, particularly anchoring to test values, including the initial test value; see Green et al. (1998) and Hurd et al. (1998). The data generation process may then be a convolution of the target distribution and a distribution of psychometric errors. This paper will ignore these issues and treat the data generation process as if it is the target distribution. However, we do empirically apply our estimators separately to first round and follow up bids, and find differences in the results, which provides evidence that such biases are present. The difficult general problem of deconvoluting a target distribution in the presence of psychometric errors is left for future research.

exposed to an environmental hazard is sacrificed for testing and y is one if an abnormality is found by the test at time v, so G is then the distribution of survival time W until the onset of abnormality. In a dosage response model, y is one if a lethal dose W exceeds a treatment dose v. For materials testing, at treatment level v, y is one if the material meets some requirement, e.g., G could be the distribution of speeds W at which a car safety device fails, with y indicating failure at test speed v.

A common procedure is to completely parameterize W, e.g., to assume W equals $X^\top\theta_0 - \varepsilon$ with $\varepsilon \sim N(\alpha_0, \sigma^2)$. The model then takes the form of a standard probit $Y = I[X^\top\theta_0 - V > \varepsilon]$ and can be estimated using maximum likelihood. However, estimation of the features of the distribution of W differs from ordinary binomial response model estimation when the model is not fully parameterized, because the goal is estimation of moments or quantiles of W, rather than response or choice probabilities of Y. So, for example, in the above parameterized model $E(W \mid X = x) = X^\top\theta_0 - \alpha_0$, and therefore any binomial response model estimator that fails to estimate the location term α_0, such as the semiparametrically efficient estimator of Klein and Spady (1993), is inadequate for estimation of moments of W.

Another important difference is the role of the support of V. By construction $G(v \mid x) = E(Y \mid V = v, X = x)$, so G can be estimated using ordinary parametric, semiparametric, or nonparametric conditional mean estimation. But nonparametric estimation of moments of W then requires identification of $G(v \mid x)$ everywhere on the support of W, so nonparametric identification requires that the support of V contain the support of W. However, virtually all experiments only consider a small number of values for v. While the literature contains many estimators of moments of W,[2] virtually all of them are parametric or semiparametric, using functional form assumptions to obtain identification, without recognizing or acknowledging the resulting failure of nonparametric identification. We show in an appendix that, given a fixed discrete design for V, even assuming that $W = m(X) - \varepsilon$ with X and ε independent is still not sufficient for identification, though identification does become possible in this case if $m(X)$ is finitely parameterized.

Our nonparametric estimators obtain identification by assuming either that bids v are draws from a continuously distributed random variable V, or that the experimental design varies with the sample size n, so for any fixed n there may be a finite number of values bids can take on, but this number of possible bid values becomes dense in the support of W as n goes to infinity.[3] We also

[2] See, e.g., Kanninen (1993) and Crooker and Herriges (2004) for comparisons of various, mostly parametric, WTP estimators. Estimators that are not fully parameterized include Chen and Randall (1997), Creel and Loomis (1997), and An (2000) for WTP and Ramgopal, Laud, and Smith (1993), and Ho and Sen (2000) for bioassay.

[3] Virtually all existing contingent valuation data sets draw bids from discrete distributions. However, large surveys typically have bid distributions with more mass points than small surveys, consistent with our assumption of an increasing number of bid values as sample size grows. See, e.g., Crooker and Herriges (2000) for a study of WTP bid designs, with explicit consideration of varying numbers of mass points.

show how this dependence of survey design on sample size affects the resulting limiting distributions, and we provide an alternative identifying assumption based on a semiparametric specification of W described below.[4]

With an estimate of G and sufficient identifying assumptions, features of the distribution of W such as moments and quantiles can be readily recovered, in particular, moments $\mu_r(x) = E(r(W, X) \mid X = x)$ for given functions r can be obtained by integrating (over the support of W) $r(w, x)$ times an estimator of $-\partial G(w \mid x)/\partial w$, the density of W. We instead provide direct semiparametric and nonparametric estimators of $\mu_r(x)$ that do not require an initial estimator of G. These estimators exploit the feature in our model that v is determined by experimental design. For example, we provide estimators of unconditional moments of W (or moments conditioned on discrete X) that do not require kernels or other smoothers, given knowledge of the experimental design, specifically, the limiting distribution of bids. For continuous X or when the bid distribution is unknown, our estimators still have the advantage over direct estimators of not requiring first stage estimation of the derivative of G.

We consider estimation for a few different information sets. In the most general case, the distribution of $W|X$ is completely unspecified apart from smoothness, and is nonparametrically estimated. We may write this case as $W = m(X, \varepsilon)$ for an unobserved ε. This includes as a special case, and is strictly weaker than, the location model $W = m(X) - \varepsilon$, where the function m and the distribution of ε are unknown.

The second case we analyze is the semiparametric model $W = \Lambda[m(X, \theta_0) - \varepsilon]$ for known functions m and Λ, where the parameters θ_0 and the distribution of $\varepsilon \perp X$ are unknown. This model includes as special cases the above probit model as well as logit and (with Λ exponential and ε extreme value) the Weibull proportional hazards model for G. In this semiparametric model, identification requires that the support of $m(X, \theta_0) - \Lambda^{-1}(V)$ become dense in the support of ε, so in this semiparametric case identification is possible with a fixed, discrete design for V, given the presence of a continuously distributed element of X.[5]

In either of these two cases (nonparametric or semiparametric W), the asymptotic design distribution of the bid values may either be known or unknown to the researcher, which yields a total of four different estimation scenarios. We provide estimators, and associated limiting normal distributions, for each of these four situations, since each is relevant for some applications. We also provide Monte

[4] An approach that we do not pursue in this paper is to sacrifice point identification and instead estimate bounds on features of G, as in McFadden (1998). See also Manski and Tamer (2002).

[5] Other possible identifying assumptions might include homogeneity as in Matzkin's (1992) threshold crossing model, or An's (2000) model which assumes W is an unknown monotonic transformation of $X^\top \theta_0 + \varepsilon$ with the distribution of ε known. See also Manski and Tamer (2002) and Das (2002).

Carlo analyses of the estimators, and an empirical application estimating conditional mean WTP to protect wetland habitats in California's San Joaquin Valley.

2 Estimators

2.1 The Data Generation Process and Estimands

Let $G(w \mid x) = \Pr(W > w \mid X = x)$, so G is the unknown complementary cumulative distribution function of a latent, continuously distributed unobserved random scalar W, conditioned on a vector of observed covariates X. Let $g(w \mid x)$ denote the conditional probability density function of W, so $g = -dG/dw$. A test value v (a realization of V) is set by an experimental design or natural experiment. Define $Y = 1(W > V)$ where $1(\cdot)$ is the indicator function. The observed data consist of a sample of realizations of covariates X, test values V, and outcomes Y. The framework is similar to random censored regressions (with censoring point V), except that for random censoring we would observe W for observations having $W > V$, whereas in the present context we only observe $Y = I(W > V)$.

Given a function $r(w, x)$, the goal is estimation of the conditional moment $\mu_r(x) = E[r(W, X) \mid X = x]$ for any chosen x in the support of X. Of particular interest are the moments based on $r(W, X) = W^k$ for integers k. We also consider estimation of quantiles. We assume the conditional distribution of W given $X = x$ is not finitely parameterized, since otherwise ordinary maximum likelihood estimation would suffice.

Assumption A.1. *The covariate vector X has support $X \subseteq \mathbb{R}^d$. The latent scalar W has an unknown, twice continuously differentiable, strictly monotonic, conditional CDF $1 - G(w \mid x)$ with probability density function $g(w \mid x)$ and a compact support $[\rho_0(x), \rho_1(x)]$. The variables W and V are conditionally independent, given X. Let $Y = I(W > V)$. Let G^{-1} be the inverse of the function G with respect to its first element.*

Assumption A.2. *The function $r(w, x)$, chosen by the researcher, is regular, meaning that it is continuous in (w, x) for all w and x on their supports, and for each x is twice continuously differentiable in w. Define $r'(w, x) = \partial r(w, x)/\partial w$. Let $\kappa(x)$ be a function or constant in $[\rho_0(x), \rho_1(x)]$. The moment $\mu_r(x)$ exists, defined by $\mu_r(x) = E[r(W, X) \mid X = x]$.*

It follows immediately from Assumption A.1, in particular the conditional independence of W and V, that

$$G(v \mid x) = E(Y \mid V = v, X = x). \tag{1}$$

and if $G(v \mid x)$ can be estimated for all $v \in \text{supp}(W)$, then conditional moments $\mu_r(x)$ could be estimated using

$$\mu_r(x) = \int_{\text{supp}(W)} r(v,x) \frac{d[1 - G(v \mid x)]}{dv} dv.$$

The disadvantage of this expression is that it involves the derivative of a high dimensional function $G(v|x)$. We apply an integration by parts to this expression to obtain the basis for more direct estimators of $\mu_r(x)$.

If $G(w \mid x)$ is not at least partly parameterized, then equation (1) implies that for identification of the distribution of W, the support of V should contain the support of W. As noted in the introduction, and by Theorem 5 in the Appendix, the distribution of W is in general not identified when the support of V has a finite number of elements. To identify features of the distribution of W with minimal restrictions on G, our nonparametric estimators assume an experimental design in which the number of mass points may grow to infinity with the sample size, as follows.

Let $H_n(v, x \mid n)$ denote the realization of the observed sample of size n, which includes both nature's selection of X and the experimental design that selects V given X. Realizations could be random draws from a CDF $H(v, x \mid n)$, but the data, particularly bids, could also be derived from some purposive sampling protocol. The requirement we place on the data generating process is the following.

ASSUMPTION A.3. *Let $H_n(v, x \mid n)$ denote the empirical CDF of V, X, for sample size n. $\sup_v |H_n(v, x \mid n) - H(v, x)| \to 0$ a.s., where $H(v, x)$ is a CDF having the property that the corresponding conditional distribution of V given $X = x$, denoted $H(v \mid x)$, has a strictly positive continuous probability density function $h(v \mid x)$ with compact support $[\delta_0(x), \delta_1(x)]$ such that $\delta_0(x) \leq \rho_0(x)$ and $\delta_1(x) \geq \rho_1(x)$.*

Assumption A.3 is used to obtain nonparametric identification. For obtaining limiting distributions it will also be assumed that $n^\tau [H_n(v, x \mid n) - H(v, x)]$ converges weakly to a Gaussian process for some τ, with $\tau = 1/2$ for root n asymptotics. Two examples illustrate this data generating process assumption:

1. Suppose for each sample observation $i = 1, \ldots, n$, X_i, V_i is drawn randomly from the CDF $H(v, x)$. Then the required sup norm convergence follows by the Glivenko-Cantelli theorem, and the convergence to a Gaussian process with $\tau = 1/2$ can be shown by, e.g., the Shorack and Wellner (1986 p. 108ff) treatment of triangular arrays of empirical processes.

2. For each sample size n, a design with J_n possible values of V is selected, and let the set of values (the support) be denoted $\mathcal{J}_n$. Suppose this design has the property that the maximum distance between a point in the support of W and a design point is of order $1/J_n$, and that

$n^{\tau-1}J_n \to \infty$. Suppose X_i is drawn randomly from a distribution, and V_i is drawn randomly from a discrete distribution $H(v \mid X_i, n)$ whose support is $\mathcal{J}_n$. Under some further conditions we can expect $n^{1/2}\left[H_n(v, x \mid n) - H(v, x \mid n)\right]$ to satisfy a triangular array functional central limit theorem, while $H(v, x \mid n) - H(v, x)$ is uniformly of order J_n^{-1}. This case covers [or would cover when the design sequence is spelled out satisfying the condition on J_n and the convergence properties of $H(v \mid X_i, n)$] all current studies, at least up to the quality of the asymptotic approximation of the design.

In our simulation studies, we will examine the size of finite sample bias that results when our estimators are applied both with discrete V and continuous V.

For estimation we suppose that a sample $Z_i = (X_i, V_i, Y_i)$ for $i = 1, \ldots, n$ is observed, generated in accordance with Assumption A.3, where V_i is a realization of V, Y_i is a realization of Y, and X_i is a realization of X. Using this data, we will provide five different estimators for $\mu_r(x)$, denoted $\widehat{\mu}_{jr}(x)$ for $j = 1, 2, 3, 4, 5$. Each is appropriate for different information sets.

The estimator $\widehat{\mu}_{1r}(x)$ is for nonparametric estimation when the limiting experimental design density $h(v \mid x)$ is known, and $\widehat{\mu}_{2r}(x)$ is for nonparametric estimation when $h(v \mid x)$ is unknown. Similarly, $\widehat{\mu}_{3r}(x)$ and $\widehat{\mu}_{4r}(x)$ cover the cases of semiparametric estimators where W is parameterized up to an unknown error term, with $h(v \mid x)$ known and unknown, respectively. An additional semiparametric estimator $\widehat{\mu}_{5r}(x)$ is provided that is simpler than $\widehat{\mu}_{3r}$ or $\widehat{\mu}_{4r}$, but may only be used for certain choices of r.

2.2 Nonparametric Moments

THEOREM 1. *Let Assumptions A.1 and A.2 hold. Let $h(v \mid x)$ be a strictly positive conditional probability density function, and $H(v \mid x)$ be the associated CDF having compact support $[\delta_0(x), \delta_1(x)]$ such that $\delta_0(x) \leq \rho_0(x)$ and $\delta_1(x) \geq \rho_1(x)$. Define*

$$s_r(x, v, y) = r[\kappa(x), x] + \frac{r'(v, x)[y - 1(v < \kappa(x))]}{h(v \mid x)}$$

$$t_r(x, v) = \frac{r'(v, x)[G(v \mid x) - 1(v < \kappa(x))]}{h(v \mid x)}.$$

Then

$$\mu_r(x) = r[\kappa(x), x] + \int_{\rho_0(x)}^{\rho_1(x)} t_r(x, v) H(dv \mid x). \tag{2}$$

Also, if V is drawn from a conditional CDF $H(v \mid x, n)$ at sample size n, then

$$\mu_r(x) = E\left[s_r(X, V, Y) \mid X = x\right] + \int_{\rho_0(x)}^{\rho_1(x)} t_r(x, v)[H(dv \mid x) - H(dv \mid x, n)] \tag{3}$$

and, if Assumption A.3 also holds, as $n \to \infty$,

$$\mu_r(x) = \lim_{n\to\infty} E\left[s_r(X,V,Y) \mid X = x\right]. \tag{4}$$

PROOF OF THEOREM 1. Starting from the definition of $\mu_r(x)$,

$$\begin{aligned}\mu_r(x) &= \int_{\rho_0(x)}^{\rho_1(x)} r(v,x)g(v \mid x)dv \\ &= \int_{\rho_0(x)}^{\kappa(x)} r(v,x)\frac{d[1-G(v \mid x)]}{dv}dv + \int_{\kappa(x)}^{\rho_1(x)} r(v,x)\frac{-dG(v \mid x)}{dv}dv\end{aligned}$$

and applying integration by parts to each of the above integrals yields

$$\begin{aligned}\mu_r(x) &= r[\kappa(x),x] + \int_{\rho_0(x)}^{\rho_1(x)} r'(v,x)[G(v \mid x) - 1(v < \kappa(x))]dv \\ &= r[\kappa(x),x] + \int_{\rho_0(x)}^{\rho_1(x)} \frac{r'(v,x)[G(v \mid x) - 1(v < \kappa(x))]}{h(v \mid x)}H(dv \mid x),\end{aligned}$$

which is equation (2). Adding and subtracting $\int_{\rho_0(x)}^{\rho_1(x)} t_r(x,v)H(dv \mid x,n)$ gives

$$\begin{aligned}\mu_r(x) = r[\kappa(x),x] + \int_{\rho_0(x)}^{\rho_1(x)} \frac{r'(v,x)[G(v \mid x) - 1(v < \kappa(x))]}{h(v \mid x)}H(dv \mid x,n) \\ + \int_{\rho_0(x)}^{\rho_1(x)} t_r(x,v)[H(dv \mid x) - H(dv \mid x,n)],\end{aligned}$$

which yields equation (3) after applying the law of iterated expectations to replace $G(v \mid x)$ by Y in the first integral. Equation (4) then follows from the convergence in Assumption A.3 and the bounded continuity of t_r. ■

We can use equation (3) to compute an estimator of $\mu_r(x)$ by the analogy principle substituting in estimators of the unknown quantities. Let $\widehat{\mu}_{1r}(x)$ denote this estimator, details supplied below. The estimator $\widehat{\mu}_{1r}(x)$ is numerically simple (and in particular does not require kernel or other smoothers if X is discrete), but requires the researcher to know, or be able to estimate, the limiting design density $h(v \mid X)$.[6] An estimator that does not entail knowing or estimating the limiting density h can be constructed as follows. First observe that equation (2) in Theorem 1 does not require Assumption A.3, so the CDF $H(v \mid x)$ and associated density $h(v \mid x)$ need not describe the limiting

[6]If h is unknown, then based on $\widehat{\mu}_{1r}$ an estimator of $\mu_r(x)$ could be constructed by first estimating h. Specifically, one could replace $h(v \mid x)$ with an estimate $\widehat{h}(v \mid x)$ (using, e.g., kernel density estimation) in the definition of $s_r(x,v,y)$. Call the result $\widehat{s}_r(x,v,y)$. The estimator of $\mu_r(x)$ would then be $\widehat{\mu}^*_{1r}(x) = \widehat{E}[\widehat{s}_r(X,V,Y) \mid X = x]$

data generating process for V, but may simply be chosen for convenience or efficiency. In particular, letting $H(v \mid x)$ be a uniform distribution reduces equation (2) to

$$\mu_r(x) = r[\kappa(x), x] + \int_{\rho_0(x)}^{\rho_1(x)} r'(v, x)[G(v \mid x) - 1(v \geq \kappa(x))]dv. \tag{5}$$

Let a_0 and a_1 be known or estimated constants such that $a_0 \leq \rho_0(x)$ and $a_1 \geq \rho_1(x)$. Then, by equations (5) and (1), a consistent estimator of $\mu_r(x)$ is given by

$$\widehat{\mu}_{2r}(x) = r[\kappa(x), x] + \int_{a_0}^{a_1} r'(v, x)[\widehat{E}(Y \mid V = v, X = x) - 1(v < \kappa)]dv, \tag{6}$$

where $\widehat{E}(Y \mid V = v, X = x)$ is an estimate of $E(Y \mid V = v, X = x)$. One could construct additional analogous estimators based on (2) instead of (5), using other choices of H, but for simplicity, we apply Theorem 1 only in the form of equations (3) and (5). Finally, one can compute

$$\widehat{\mu}_{0r}(x) = \mu_r(x) = -\int_{\text{supp}(W)} r(v, x)\widehat{g}(v \mid x)dv,$$

where $\widehat{g}(v \mid x)$ is an estimate of $g(v \mid x) = \partial G(v \mid x)/\partial v$.

Consistency and potential effects of finite sample design on limit distributions for $\widehat{\mu}_{2r}(x)$ are analogous to the above discussion of $\widehat{\mu}_{1r}(x)$. In applications, the choice between using $\widehat{\mu}_{1r}(x)$ or $\widehat{\mu}_{2r}(x)$ would be based at least in part on the information set of the researcher regarding the limiting design density. We provide more details later on the construction and limiting distributions of these estimators.

In the special case of the nonparametric location model $W = \Lambda[m(X) - \varepsilon]$ with $\varepsilon \perp X$, and Λ known and invertible, these $\mu_r(x)$ estimators can be used to estimate an unknown $m(x)$, since $m(x) = \mu_r(x) - E(\varepsilon)$ with $r(w, x) = \Lambda^{-1}(w)$. Chen and Randall (1997) and Crooker and Herriges (2004) consider this case. An (2000) considers the model where Λ is unknown but m and the distribution of ε are known; this also is a special case of our nonparametric model.

We summarize the three estimators below in terms of the quantities they assume are known and in terms of their 'curse factor' - this is a vector indicating the highest dimensions d of nonparametric estimation carried out along with the highest degree of derivative estimated ν. According to Stone (1980) the optimal pointwise rate for estimating regression functions and their derivatives is of order $n^{-(p-\nu)/(d+2p)}$, where p is the degree of smoothness of the function. In our case this difficulty is washed out to first order in the asymptotics but is present in higher order terms, making estimators with higher curse factor less attractive.

Estimator	$\widehat{\mu}_{0r}(x)$	$\widehat{\mu}_{1r}(x)$	$\widehat{\mu}_{2r}(x)$
Known	-	$h(v \mid x)$	-
Curse Factor[7]	$(d+1, 1)$	$(d, 0)$	$(d+1, 0)$

[7]The vector (d, p) gives the highest dimensional nonparametric estimation (d) that is used and the highest number

Note that we can estimate unconditional moments $\mu_r = E[r(W, X)]$ at rate root n without any smoothing and without any parameter estimation, using $E[s_r(X, V, Y)]$.

2.3 Semiparametric Moments

Corollary 1 below will be used in place of Theorem 1 to obtain faster convergence rates using a semiparametric model for W.

ASSUMPTION A.4. *The latent W satisfies $W = \Lambda[m(X, \theta_0) - \varepsilon]$, where m and Λ are known functions, Λ is invertible and differentiable with derivative denoted Λ', $\theta_0 \in \Theta$ is a vector of parameters, and ε is a disturbance that is distributed independently of V, X, with unknown, twice continuously differentiable CDF $F_\varepsilon(\varepsilon)$ and compact support $[a_0, a_1]$ that contains zero. Define $U = m(X, \theta_0) - \Lambda^{-1}(V)$. Let $\Psi_n(U \mid n)$ denote the empirical CDF of U at sample size n. $\sup_v |\Psi_n(U \mid n) - \Psi(U)| \to 0$ a.s., where $\Psi(U)$ is a CDF that has an associated PDF $\psi(U)$ that is continuous and strictly positive on the interval $[a_0, a_1]$.*

Define $s_r^*(x, u, y)$ and $t_r^*(x, u)$ by

$$s_r^*(x, u, y) = r[\Lambda(m(x, \theta_0)), x] + \frac{r'[\Lambda(m(x, \theta_0) - u), x]\Lambda'(m(x, \theta_0) - u)[y - 1(u > 0)]}{\psi(u)}.$$

$$t_r^*(x, u) = \frac{r'[\Lambda(m(x, \theta_0) - u), x]\Lambda'(m(x, \theta_0) - u)[F_\varepsilon(u) - 1(u > 0)]}{\psi(u)}.$$

If Λ is the identity function, then W equals a parameterized function of x plus an additive independent error. If Λ is the exponential function, then it is $\ln(W)$ that is modeled with an additive error. Other examples include: the Box-Cox, $\Lambda^{-1}(W) = (W^\lambda - 1)/\lambda$, the Zellner-Revankar $\Lambda^{-1}(W) = \ln W + \lambda W$, and the arcsinh $\Lambda^{-1}(W) = \sinh^{-1}(\lambda W)/\lambda$, where in each case λ is a free parameter.

COROLLARY 1. *Let Assumptions A.1, A.2, and A.4 hold. Then*

$$E(Y \mid U = u) = F_\varepsilon(u)$$

$$\mu_r(x) = r[\Lambda(m(x, \theta_0)), x] + \int_{a_0}^{a_1} t_r^*(x, u)\Psi(du),$$

$$\mu_r(x) = E\left[s_r^*(x, U, Y)\right] + \int_{a_0}^{a_1} t_r^*(x, u)[\Psi(du) - \Psi(du \mid n)] = \lim_{n\to\infty} E\left[s_r^*(x, U, Y)\right]$$

and, if Assumption A.3 also holds,

$$\Psi_n(u \mid n) = E(1 - H[\Lambda(m(X, \theta_0) - u) \mid X, n])$$

of derivatives (p) that are estimated.

$$\psi_n(u) = E\left[h[\Lambda(m(X,\theta_0)-u) \mid X]\Lambda'(m(X,\theta_0)-u)\right] \to \psi(u)$$

PROOF OF COROLLARY 1. Recall that $Y = I(W > V) = I(\varepsilon < U)$, so $E(Y \mid U = u) = F_\varepsilon(u)$. Starting from the definition of $\mu_r(x)$,

$$\begin{aligned}\mu_r(x) &= \int_{a_0}^{a_1} r[\Lambda(m(x,\theta_0)-\varepsilon), x]F_\varepsilon(d\varepsilon) \\ &= \int_{a_0}^{0} r[\Lambda(m(x,\theta_0)-u), x]\frac{dF_\varepsilon(u)}{du}du + \int_{0}^{a_1} r[\Lambda(m(x,\theta_0)-u), x]\frac{d[F_\varepsilon(u)-1]}{du}du\end{aligned}$$

and applying integration by parts to each of the above integrals yields

$$\begin{aligned}\mu_r(x) &= r[\Lambda(m(x,\theta_0)), x] + \int_{a_0}^{a_1} r'[\Lambda(m(x,\theta_0)-u), x]\Lambda'(m(x,\theta_0)-u)[F_\varepsilon(u) - I(u>0)]du \\ &= r[\Lambda(m(x,\theta_0)), x] + \int_{a_0}^{a_1} t_r^*(x,u)\Psi(du) \\ &= r[\Lambda(m(x,\theta_0)), x] + \int_{a_0}^{a_1} t_r^*(x,u)\Psi_n(du \mid n) + \int_{a_0}^{a_1} t_r^*(x,u)[\Psi(du) - \Psi_n(du \mid n)]\end{aligned}$$

Next, apply the law of iterated expectations to obtain

$$\begin{aligned}E\left[s_r^*(X,U,Y)\right] &= E\left(\frac{r'[\Lambda(m(x,\theta_0)-u), x]\Lambda'(m(x,\theta_0)-u)[F_\varepsilon(u) - 1(u>0)]}{\psi(u)}\right) \\ &= \int_{a_0}^{a_1} t_r^*(x,u)\Psi_n(du \mid n),\end{aligned}$$

which gives the expressions for $\mu_r(x)$, and $\int_{a_0}^{a_1} t_r^*(x,u)[\Psi(du) - \Psi_n(du \mid n)] \to_p 0$ by the uniform convergence of Ψ_n.

Note that $\Psi_n(u \mid n)$ is the empirical probability that $U \leq u$, which is the same event as $V \geq \Lambda(m(X,\theta_0) - u)$. Conditioning on $X = x$ this probability would be $1 - H_n[\Lambda(m(x,\theta_0) - u) \mid x, n]$, and averaging over X gives $\Psi_n(u \mid n) = E(1 - H_n[\Lambda(m(X,\theta_0) - u) \mid X, n])$. This implies $\Psi(u) = \lim_{n\to\infty} E(1 - H[\Lambda(m(X,\theta_0) - u) \mid X])$, where the only role of the limit is to evaluate the expectation at the limiting distribution of X. Taking the derivative with respect to u gives $\psi(u) = \lim_{n\to\infty} E(h[\Lambda(m(X,\theta_0) - u \mid X)\Lambda'(m(X,\theta_0) - u)])$. Consistency of $\psi_n(u)$ then follows from the uniform convergence of the distribution of X to its limiting distribution in Assumption A.3. ■

Now consider rate root n estimation of arbitrary conditional moments based on Corollary 1. It will be convenient to first consider the case where θ_0 is known, implying that the conditional mean of W is known up to an arbitrary location (since ε is not required to have mean zero). A special case of known θ_0 is when x is empty, i.e., estimation of unconditional moments of W, since in that case we can without loss of generality take m to equal zero.

2.3.1 Estimation With Known θ

Suppose that θ_0 is known. Considering first the case where the limiting design density $h(v|x)$ is also known, for a given u define the sample average $\widehat{\psi}(u)$ by

$$\widehat{\psi}(u) = \frac{1}{n}\sum_{i=1}^{n} h[\Lambda(m(X_i,\theta_0) - u) \mid X_i]\Lambda'(m(X_i,\theta_0) - u).$$

Then, based on Corollary 1, we have consistency of the estimator

$$\widehat{\mu}^*_{3r}(x) = r[\Lambda(m(x,\theta_0)), x] + \frac{1}{n}\sum_{i=1}^{n} \frac{r'[\Lambda(m(x,\theta_0) - U_i), x]\Lambda'(m(x,\theta_0) - U_i)[Y_i - 1(U_i > 0)]}{\widehat{\psi}(U_i)}.$$

This estimator is computationally extremely simple, since it entails only sample averages. Special cases of this estimator were proposed by McFadden (1994) and by Lewbel (1997).

Let $\widetilde{\psi}(u)$ be an estimator of $\psi(u)$ that does not depend on knowledge of h. For example $\widetilde{\psi}(u)$ could be a (one dimensional) kernel density estimator of the density of U, based on the data $\widehat{U}_i$ and evaluated at u. We then have the estimator

$$\widehat{\mu}^*_{4r}(x) = r[\Lambda(m(x,\theta_0)), x] + \frac{1}{n}\sum_{i=1}^{n} \frac{r'[\Lambda(m(x,\theta_0) - U_i), x]\Lambda'(m(x,\theta_0) - U_i)[Y_i - 1(U_i > 0)]}{\widetilde{\psi}(U_i)},$$

which may be used when h is unknown.

Finally, similarly to the nonparametric estimator $\widehat{\mu}_{0r}(x)$ one can compute

$$\widehat{\mu}^*_{5r}(x) = \int_{a_0}^{a_1} r[\Lambda(m(x,\theta_0) - \varepsilon), x]\widehat{f}_\varepsilon(\varepsilon)d\varepsilon, \quad \widehat{f}_\varepsilon(\varepsilon) = \left.\frac{d\widehat{E}(Y \mid U)}{dU}\right|_{U=\varepsilon},$$

where $\widehat{E}(Y \mid U)$ is a nonparametric estimator of $E(Y \mid U)$ based on $\{Y_i, U_i\}_{i=1}^n$ with $U_i = m(X_i,\theta_0) - \Lambda^{-1}(V_i)$.

2.3.2 Estimation of θ

First, consider estimation of θ. By Assumption A.4,

$$E[\Lambda^{-1}(W) \mid X = x] = \alpha_0 + m(x,\theta_0)$$

for some arbitrary location constant α_0. This constant is unknown since no location constraint is imposed upon ε. Let $s_{\Lambda^{-1}}(X,V,Y)$ denote $s_r(X,V,Y)$ with $r(w,x) = \Lambda^{-1}(w)$. It then follows from Theorem 1 that

$$\lim_{n\to\infty} E\left[s_{\Lambda^{-1}}(X,V,Y) \mid X = x\right] = \lim_{n\to\infty} E(\Lambda^{-1}(W) \mid X = x).$$

Note that the limit as $n \to \infty$ means that the expectations are taken at the limiting distributions of the data. In other words the asymptotic conditional expectation of the known or estimable quantity $s_{\Lambda^{-1}}$ is equal to $\alpha_0 + m(x, \theta_0)$. Under some identification conditions this can be used for estimation of (α_0, θ_0). Specifically, we could estimate θ_0 by minimizing the least squares criterion

$$(\widehat{\theta}, \widehat{\alpha}) = \arg\min_{\theta,\alpha} \frac{1}{n} \sum_{i=1}^{n} [s_{\Lambda^{-1}}(X_i, V_i, Y_i) - \alpha - m(X_i, \theta)]^2. \tag{7}$$

If m is linear in parameters, then a closed form expression results for both parameter estimates. If h is not known, one could replace $h(V \mid X)$ in the expression of $s_{\Lambda^{-1}}(X, V, Y)$ with an estimate $\widehat{h}(V \mid X)$. The resulting estimator would then take the form of a two step estimator with a nonparametric first step (the estimation of h). This estimator of θ and α is equivalent to the estimator for general binary choice models proposed by Lewbel (2000), though Lewbel provides other extensions, such as to estimation with endogenous regressors.

With Assumption A.4, the latent error ε is independent of X, and therefore the binary choice estimator of Klein and Spady (1993) may provide a semiparametrically efficient estimator of θ.[8]

2.3.3 Estimation with Unknown θ

Let $\widehat{\theta}$ denote a root n consistent, asymptotically normal estimator for θ_0. Replacing θ_0 with any $\theta \in \Theta$ we may rewrite the estimators of the previous section as $\widehat{\mu}^*_{\lambda r}(x; \theta)$ for $\lambda = 3$, or 4. In doing so, note that θ appears both directly in the equations for $\widehat{\mu}^*_{\lambda r}$, and also in the definition of $U_i = m(X_i, \theta) - \Lambda^{-1}(V_i)$. We later derive the root n consistent, asymptotically normal limiting distribution for each estimator $\widehat{\mu}_{\lambda r}(x) = \widehat{\mu}^*_{\lambda r}(x; \widehat{\theta})$, where we suppress the dependence on $\widehat{\theta}$ for simplicity. The estimators are not differentiable in U_i (except $\widehat{\mu}^*_{5r}(x)$)[9], which complicates the derivation of their limiting distribution, e.g., even with a fixed design, Theorem 6.1 of Newey and McFadden (1994) is not be directly applicable due to this nondifferentiability.

We summarize the three estimators below in terms of the quantities they assume are known and in terms of their 'curse factor'.

Estimator	$\widehat{\mu}_{3r}(x)$	$\widehat{\mu}_{4r}(x)$	$\widehat{\mu}_{5r}(x)$
Known	Λ, m, h	Λ, m	Λ, m
Curse Factor	$(0, 0)$	$(1, 0)$	$(1, 1)$

[8] The Klein and Spady estimator does not identify a location constant α, but that is not required for this step, since no location constraint is imposed upon ε. Also, for the present application, the limiting distribution theory for Klein and Spady would need to be extended to allow for data generating processes that vary with the sample size.

[9] But this estimator depends on nonparametric estimates of the derivative of a regression function. This estimator shares the asymptotics of weighted average derivative estimators.

3 Estimation Details and Distribution Theory

In this section we provide more detail about the computation of the estimators $\widehat{\mu}_{1r}(x), \ldots, \widehat{\mu}_{5r}(x)$ and their distribution theory.

3.1 Nonparametric Estimators

There are many different nonparametric methods for estimating regression functions. For purely continuous variables with density bounded away from zero throughout their support the local linear kernel method is attractive. This method has been extensively analyzed and has some positive properties like being design adaptive, and best linear minimax under standard conditions; see Fan and Gijbels (1996) for further discussion.[10] One issue we are particularly concerned about is how to handle discrete variables. Specifically, some elements of X could be discrete, either ordered discrete or unordered discrete, while V can be ordered discrete. When there is a single discrete variable that takes only a small number of values, the pure frequency estimator is the natural and indeed optimal estimator to take in the absence of additional structure. In fact, one obtains parametric rates of convergence in the pure discrete case [and in the mixed discrete/continuous case the rate of consistency is unaffected by how many such discrete covariates there are], see Delgado and Mora (1995) for discussion. When there are many discrete covariates, it may be desirable to use some 'discrete smoothing', as discussed in Li and Racine (2002), see also Wang and Van Ryzin (1981). Coppejans (2003) considers a case most similar to our own - he allows the distribution of the discrete data to change with sample size. One major difference is that his data have arrived from a very specific grouping scheme that introduces an extra bias problem.

We shall not outline all the possibilities for estimation here with regard to the covariates X, rather we assume that X is continuously distributed with density bounded away from zero. However, the estimators we define can be applied in all of the above situations [although they may not be optimal], and the estimators are still asymptotically normal with the rate determined by the number of continuous variables.

We will pay more attention to the potential discreteness in V, since this is key to our estimation problem. For clarity we will avoid excessive subscripts/superscripts. We suppose that V is asymptotically continuous in the sense that for each n, V_i is drawn from a distribution $H(v|X_i, n)$ that has

[10] If there is a continuous density but with some points in the support of zero density, the rate of convergence may be slower but Hengartner and Linton (1996) have shown that the local linear estimator can still achieve the optimal rate in this case. There are other non-standard cases: Lu (2002) considers the case where the covariate process has fractal dimension [e.g., in the multivariate case where the covariates lie on a nonlinear manifold of lower local dimension].

finite support, increasing with n.[11]

Under our conditions there is a bias in the estimates of $\mu_r(x)$ of order J^{-1} in this discrete case. Therefore, for this term not to matter in the limiting distribution we require that $\delta_n J^{-1} \to 0$, where δ_n is the rate of convergence of the estimator in question [$\delta_n = \sqrt{n}$ in the parametric case but $\delta_n = \sqrt{nb^d}$ for some bandwidth b in the nonparametric cases]. In the nonparametric case, the spacing of the discrete covariates is closer than the bandwidth of a standard kernel estimator, that is, we know that $b^2 J \to \infty$ so that J^{-1} is much smaller than the smoothing window of a kernel estimator. Therefore, the pure frequency estimator is dominated by a smoothing estimator, and we shall just construct smoothing-based estimators.

The estimator $\widehat{\mu}_{1r}(x)$ involves smoothing the data

$$s_r(Z_i) = r[\kappa(X_i), X_i] + \frac{r'(V_i, X_i)[Y_i - 1(V_i < \kappa(X_i))]}{h(V_i \mid X_i)}$$

against X_i, where $Z_i = (V_i, X_i, Y_i)$. Define the $p-1$-th order local polynomial regression of $s_r(Z_i)$ on X_i by minimizing

$$Q^s_{p-1,n}(\vartheta) = \frac{1}{n}\sum_{i=1}^{n} K_b(X_i - x)\left[s_r(Z_i) - \sum_{0 \le |\mathbf{j}| \le p-1} \vartheta_{\mathbf{j}} (X_i - x)^{\mathbf{j}}\right]^2 \tag{8}$$

with respect to the vector ϑ containing all the $\vartheta_{\mathbf{j}}$, where $K_b(t) = \prod_{j=1}^{d} k_b(t_j)$ with $k_b(u) = k(u/b)/b$, where k is a univariate kernel function and $b = b(n)$ is a bandwidth. Here, we are using the multi-dimensional index notation, for vectors $\mathbf{j} = (j_1, \ldots, j_d)^\top$ and $a = (a_1, \ldots, a_d)^\top$: $\mathbf{j}! = j_1! \times \cdots \times j_d!$, $|\mathbf{j}| = \sum_{k=1}^{d} j_k$, $a^{\mathbf{j}} = a_1^{j_1} \times \ldots \times a_d^{j_d}$, and $\sum_{0 \le |\mathbf{j}| \le p-1}$ denotes the sum over all $\mathbf{j}$ with $0 \le |\mathbf{j}| \le p-1$. Let $\widehat{\vartheta}_0$ denote the first element of the vector $\widehat{\vartheta}$ that minimizes (8). Then let

$$\widehat{\mu}_{1r}(x) = \widehat{\vartheta}_0. \tag{9}$$

This estimator is linear in the dependent variable and has an explicit form.

In computing the estimator $\widehat{\mu}_{2r}(x)$ we require an estimator of $G(v \mid x)$, which is given by the smooth of Y_i on X_i, V_i. Let $\widetilde{X}_i = \left(V_i, X_i^\top\right)^\top$ and $\widetilde{x} = (v, x)^\top$ and define the $p-1$-th order local polynomial regression of Y_i on $\widetilde{X}_i$ by minimizing

$$Q^Y_{p,n}(\vartheta) = \frac{1}{n}\sum_{i=1}^{n} \widetilde{K}_b(\widetilde{X}_i - \widetilde{x})\left[Y_i - \sum_{0 \le |\mathbf{j}| \le p-1} \vartheta_{\mathbf{j}} \left(\widetilde{X}_i - \widetilde{x}\right)^{\mathbf{j}}\right]^2, \tag{10}$$

[11] The case where V_i is drawn from a continuous distribution $H(v|X_i)$ for all n is really a special case of our set-up.

where $\widetilde{K}_b(\widetilde{X}_i - \widetilde{x}) = k_b(V_i - v) K_b(X_i - x)$. Let $\widehat{\vartheta}_0$ denote the first element of the vector $\widehat{\vartheta}$ that minimizes (10), and let $\widehat{G}(v \mid x) = \widehat{\vartheta}_0$. Then define

$$\widehat{\mu}_{2r}(x) = r(\kappa(x), x) + \int_{\rho_0(x)}^{\rho_1(x)} r'(v, x)[\widehat{G}(v \mid x) - 1(v < \kappa(x))]dv, \tag{11}$$

where the univariate integral is interpreted in the Lebesgue Stieltjes sense (actually under our conditions $\widehat{G}(v \mid x)$ is a continuous function and $1(v < \kappa(x))$ is a simple step function).

Finally, to compute $\widehat{\mu}_{0r}(x)$ we use one higher order of polynomial, i.e., minimize $Q_{p,n}^{Y}(\vartheta)$ with respect to ϑ, and let $\partial\widehat{G}(v \mid x)/\partial v = \widehat{\vartheta}_v$, where $\widehat{\vartheta}_v$ is the second element of the vector $\widehat{\vartheta}$. Then define

$$\widehat{\mu}_{0r}(x) = -\int_{\rho_0(x)}^{\rho_1(x)} r(v, x) \frac{\partial\widehat{G}(v \mid x)}{\partial v} dv. \tag{12}$$

When there are some discrete components to X, it may be advantageous to modify the kernel window along the lines discussed in Li and Racine (2002).

The estimator (11) is in the class of marginal integration/partial mean estimators sometimes used for estimating additive nonparametric regression models, see Linton and Nielsen (1995), Newey (1994), and Tjø stheim and Auestad (1994), except that the integrand is not just a regression function and the integrating measure λ, where (asymptotically) $d\lambda(v) = r'(v, x)1(\rho_0(x) \leq v \leq \rho_1(x))dv$, is not necessarily a probability measure, i.e., it may not be positive or integrate to one. The distribution theory for the class of marginal integration estimators has already been worked out for a number of specific smoothing methods, see the above references.

We make the following assumptions.

ASSUMPTION B.1. *k is a symmetric probability density with bounded support, and is continuously differentiable on its support.*

ASSUMPTION B.2. *The random variables (V_i, X) are asymptotically continuously distributed, i.e., for some finite constant c_h*

$$\sup_{v \in [\rho_0(x), \rho_1(x)]} |H(v|x, n) - H(v|x)| \leq \frac{c_h}{J}, \tag{13}$$

where $H(v, x)$ possesses a Lebesgue density $h(v, x)$ along with conditionals $h(v|x)$ and marginal $h(x)$. Furthermore, $\inf_{\rho_0(x) \leq v \leq \rho_1(x)} h(v, x) > 0$. For all n larger than some n_0, $\mathrm{var}(Y_i \mid V_i = v, X_i = x) < \infty$, and the limiting conditional variance $\sigma^2(v, x) = \mathrm{var}(Y_i \mid V_i = v, X_i = x) = G(v \mid x)[1 - G(v \mid x)]$. Furthermore, $G(v \mid x)$ and $h(v, x)$ are p-times continuously differentiable for all v with $\rho_0(x) \leq v \leq \rho_1(x)$, letting $g(v \mid x) = \partial G(v|x)/\partial v$ denote the conditional density of $W|X$. The set $[\rho_0(x), \rho_1(x)] \times \{x\}$ is strictly contained in the support of (V, X) for large enough n.

The condition (13) is satisfied provided the associated frequency function $h(v \mid x, n)$ satisfies $\min_{v \in \mathcal{J}_n} h(v \mid x, n) \geq \underline{v}/J_n$ and $\max_{v \in \mathcal{J}_n} h(v \mid x, n) \leq \overline{v}/J_n$ for some bounds $\underline{v} > 0$ and $\overline{v} < \infty$, and

provided the support $\mathcal{J}_n$ becomes dense in $[\rho_0(x), \rho_1(x)]$. The other conditions are standard regularity conditions for nonparametric estimation.

For a function $f : \mathbb{R}^s \to \mathbb{R}$, arrange the elements of its partial derivatives (for all vectors $\pi = (\pi_1, \ldots, \pi_s)$ such that $\sum_{j=1}^s \pi_j = p$), $\partial^{\sum_{j=1}^s \pi_j} f(t)/\partial t_1^{\pi_1} \cdots \partial t_s^{\pi_s}$ as a large column vector $f^{(p;s)}(t)$ of dimensions $(s+p-1)!/s!(p-1)!$. Let $a_{d;p}(k)$, $a_{d+1;p}(k)$ and $a^*_{d+1;p+1}(k)$ be conformable vectors of constants depending only on the kernel k, and let $c_{d;p}(k)$, $c_{d+1;p}(k)$ and $c^*_{d+1;p+1}(k)$ also be scalar kernel constants.. Define

$$\beta_0(x) = a^*_{d+1;p+1}(k)^\top \int_{\rho_0(x)}^{\rho_1(x)} r(v,x) G^{(p+1;d+1)}(v|x) dv$$

$$\beta_1(x) = a_{d;p}(k)^\top \mu_r^{(p;d)}(x) \; ; \; \beta_2(x) = a_{d+1;p}(k)^\top \int_{\rho_0(x)}^{\rho_1(x)} r'(v,x) G^{(p;d+1)}(v|x) dv,$$

$$\omega_0(x) = c^*_{d+1;p+1}(k) \int_{\rho_0(x)}^{\rho_1(x)} \sigma^2(v,x) \left(\frac{r'(v,x)h(v,x) - r(v,x)h'(v,x)}{h^2(v,x)} \right)^2 h(v,x) dv$$

$$\omega_1(x) = c_{d;p}(k) \frac{\mathrm{var}[s_r(Z) \mid X = x]}{h(x)} \; ; \; \omega_2(x) = c_{d+1;p}(k) \int_{\rho_0(x)}^{\rho_1(x)} \sigma^2(v,x) \left(\frac{r'(v,x)}{h(v,x)} \right)^2 h(v,x) dv.$$

THEOREM 2. *Suppose that assumptions A1-A3, B1 and B2 hold and that the bandwidth sequence $b = b(n)$ satisfies $b \to 0$, $nb^{d+2}/\log n \to \infty$, and $Jb^2 \to \infty$. Then, for $j = 1, 2$,*

$$\sqrt{nb^d} \left[\widehat{\mu}_{jr}(x) - \mu_r(x) - b^p \beta_j(x) \right] \Longrightarrow N(0, \omega_j(x)).$$

If G is $p+1$-times continuously differentiable, then

$$\sqrt{nb^d} \left[\widehat{\mu}_{0r}(x) - \mu_r(x) - b^p \beta_0(x) \right] \Longrightarrow N(0, \omega_0(x)).$$

REMARKS.

1. The estimator $\widehat{\mu}_{0r}(x)$ requires one more derivative than $\widehat{\mu}_{1r}(x), \widehat{\mu}_{2r}(x)$ for the results stated here.

2. In the local linear case the kernel constants in $\omega_1(x)$ and $\omega_2(x)$ are identical. A simple argument then shows that $\omega_1(x) \geq \omega_2(x)$. By the law of iterated expectation

$$\mathrm{var}[s_r(Z) \mid X = x] = E[\mathrm{var}[s_r(Z) \mid V, X] \mid X = x] + \mathrm{var}[E[s_r(Z) \mid V, X] \mid X = x].$$

Furthermore,

$$\begin{aligned} E[\mathrm{var}[s_r(Z) \quad | \quad V, X] \mid X = x] &= \int_{\rho_0(x)}^{\rho_1(x)} \left(\frac{r'(v,x)}{h(v \mid x)} \right)^2 \sigma^2(v,x) h(v \mid x) dv \\ &= h(x) \int_{\rho_0(x)}^{\rho_1(x)} \left(\frac{r'(v,x)}{h(v,x)} \right)^2 \sigma^2(v,x) h(v,x) dv. \end{aligned}$$

It follows that $\omega_1(x) \geq \omega_2(x)$. In the special case that $h'(v,x) = 0$, $\omega_0(x)$ are the same $\omega_2(x)$ apart from the kernel constants, but otherwise the ranking could go either way.

3. Regarding the biases, in the special case of local linear estimation and two continuous derivatives:

$$\begin{aligned}\beta_1(x) &\propto \sum_{j=1}^{d} \int_{\rho_0(x)}^{\rho_1(x)} \frac{\partial^2\{r(v,x)g(v \mid x)\}}{\partial x_j^2} dv \\ \beta_2(x) &\propto = \int_{\rho_0(x)}^{\rho_1(x)} \left[\sum_{j=1}^{d} \frac{\partial^2 G(v \mid x)}{\partial x_j^2} + \frac{\partial^2 G(v \mid x)}{\partial v^2}\right] r'(v,x) dv.\end{aligned}$$

Under certain conditions these two biases are the same applying integration by parts.

4. If $r(v,x)$ is a vector of functions, then the results are as above with the square operation replaced by outer product of corresponding vectors. Suppose one wants to estimate $\text{var}(W|X = x) = \mu_{W^2}(x) - \mu_W^2(x)$, a nonlinear function of the vector $(E(W^2|X=x), E(W|X=x))$. In this case, one obtains the asymptotic distribution by the delta method applied to the joint limiting behaviour of the estimators of $\mu_{W^2}(x), \mu_W(x)$.

3.2 Semiparametric Estimators

In this section we assume the conditions of A4 prevail. In this case, discreteness of V_i is less of an issue - even if V_i is discrete, if there are continuous variables in X_i, then $U_i = m(X_i, \theta_0) - \Lambda^{-1}(V_i)$ can be continuously distributed. For simplicity we therefore assume a fixed design for our limiting distribution calculations. Similar asymptotics will result when the assumption that V_i is continuously distributed is replaced by an assumption like equation (13).

Let $\widehat{\theta}$ be some consistent estimator of θ_0. Define:

$$\widehat{\mu}_{3r}(x) = r[\Lambda(m(x,\widehat{\theta})), x] + \frac{1}{n}\sum_{i=1}^{n} \frac{r'[\Lambda(m(x,\widehat{\theta}) - \widehat{U}_i), x]\Lambda'(m(x,\widehat{\theta}) - \widehat{U}_i)[Y_i - 1(\widehat{U}_i > 0)]}{\widehat{\psi}(\widehat{U}_i)}$$

$$\widehat{\mu}_{4r}(x) = r[\Lambda(m(x,\widehat{\theta})), x] + \frac{1}{n}\sum_{i=1}^{n} \frac{r'[\Lambda(m(x,\widehat{\theta}) - \widehat{U}_i), x]\Lambda'(m(x,\widehat{\theta}) - \widehat{U}_i)[Y_i - 1(\widehat{U}_i > 0)]}{\widetilde{\psi}(\widehat{U}_i)},$$

where $\widehat{U}_i = m(X_i, \widehat{\theta}) - \Lambda^{-1}(V_i)$ and

$$\widehat{\psi}(\widehat{U}_i) = \frac{1}{n}\sum_{j=1}^{n} h[\Lambda(m(X_j,\widehat{\theta}) - \widehat{U}_i)|X_j]\Lambda'(m(X_j,\widehat{\theta}) - \widehat{U}_i) \quad ; \quad \widetilde{\psi}(\widehat{U}_i) = \frac{1}{nb}\sum_{j=1}^{n} k\left(\frac{\widehat{U}_i - \widehat{U}_j}{b}\right).$$

Define also the estimators $\widehat{\mu}^*_{3r}(x)$ and $\widehat{\mu}^*_{4r}(x)$ as the special cases of $\widehat{\mu}_{3r}(x)$ and $\widehat{\mu}_{4r}(x)$ in which θ is known, in which case $\widehat{U}_i$ is replaced by U_i.

We next state the asymptotic properties of the conditional moment estimators based on Corollary 1. We need some conditions on the estimator and on the regression functions and densities.

ASSUMPTION C.1. *Suppose that*

$$\sqrt{n}(\widehat{\theta} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varsigma(Z_i, \theta_0) + o_p(1)$$

for some function ς *such that* $E[\varsigma(Z_i, \theta_0)] = 0$ *and* $\Omega = E[\varsigma(Z_i, \theta_0)\varsigma(Z_i, \theta_0)^\top] < \infty$. *Suppose also that* θ_0 *is an interior point of the parameter space.*

ASSUMPTION C.2. *The function* m *is twice continuously differentiable in* θ *and*

$$\sup_{\|\theta-\theta_0\|\leq\delta_n} \left\| \frac{\partial m}{\partial \theta}(x, \theta) \right\| \leq d_1(x) \quad ; \quad \sup_{\|\theta-\theta_0\|\leq\delta_n} \left\| \frac{\partial^2 m}{\partial \theta \partial \theta^\top}(x, \theta) \right\| \leq d_2(x)$$

with $Ed_1^r(X_i) < \infty$ *and* $Ed_2^r(X_i) < \infty$ *for some* $r > 2$.

ASSUMPTION C.3. *The density function* h *is continuous and is strictly positive on its compact support and is twice continuously differentiable. The transformation* Λ *is three times continuously differentiable.*

ASSUMPTION C.4. *The kernel* k *is twice continuously differentiable on its support, and therefore* $\sup_t |k''(t)| < \infty$. *The bandwidth* b *satisfies* $b \to 0$ *and* $nb^6 \to \infty$.

The regularity conditions are quite standard. Assumption C4 is used for $\widehat{\mu}_{4r}(x)$, which is based on a one-dimensional kernel density estimator.

For each $\theta \in \Theta$ and $x \in \mathcal{X}$, define the stochastic processes:

$$f_0(Z_i, \theta) = \frac{r'[\Lambda(m(x,\theta) - U_i(\theta)), x]\Lambda'(m(x,\theta) - U_i(\theta))[Y_i - 1(U_i(\theta) > 0)]}{\psi(U_i)}$$

$$f_1(Z_i, \theta) = r[\Lambda(m(x,\theta)), x] + \frac{r'[\Lambda(m(x,\theta) - U_i(\theta)), x]\Lambda'(m(x,\theta) - U_i(\theta))[Y_i - 1(U_i(\theta) > 0)]}{\psi(U_i)}$$

where $U_i(\theta) = m(X_i, \theta) - \Lambda^{-1}(V_i)$. Then

$$\Gamma_F = \left(\frac{\partial}{\partial \theta} E\left[f_1(Z_i, \theta) \right] \right) \Big|_{\theta=\theta_0}$$

$$\Psi_F = E\left[f_0(Z_i, \theta_0) \frac{\psi'(U_i)}{\psi(U_i)} \widetilde{\gamma}_i \right] + E\left[\frac{f_0(Z_i, \theta_0)}{\psi(U_i)} \widetilde{\zeta}_{ij} \widetilde{\gamma}_j \right]$$

$$\widetilde{\gamma}_i = \frac{\partial m}{\partial \theta^\top}(X_i, \theta_0) - E\left[\frac{\partial m}{\partial \theta^\top}(X_i, \theta_0) \right]$$

and $\zeta_{ij} = \left[h'(\Lambda|X_j)(\Lambda')^2 + h(\Lambda|X_j)\Lambda'' \right] (m(X_j, \theta_0) - U_i)$, where $\widetilde{\zeta}_{ij} = \zeta_{ij} - E_i\zeta_{ij}$.

The above quantities may depend on x but we have suppressed this notationally. Note also that $Ef_1(Z_i, \theta_0) = \mu_r(x)$.

THEOREM 3. *Suppose that Assumptions A1-A4 and C1-C3 hold. Then, as* $n \to \infty$,

$$\sqrt{n}[\widehat{\mu}_{3r}(x) - \mu_r(x)] \Longrightarrow N(0, \sigma_\eta^2(x)), \tag{14}$$

where $0 < \sigma_\eta^2(x) = \text{var}(\eta_j) < \infty$ *with* $\eta_j = \eta_{1j} + \eta_{2j} + \eta_{3j}$, *where:*

$$\eta_{1j} = f_0(Z_j, \theta_0) - Ef_0(Z_j, \theta_0)$$

$$\eta_{2j} = (\Gamma_F - \Psi_F)\, \varsigma(Z_j; \theta_0)$$

$$\eta_{3j} = -E\left[f_0(Z_i, \theta_0) \frac{h[\Lambda(m(X_j, \theta_0) - U_i)|X_j]\Lambda'(m(X_j, \theta_0) - U_i) - \psi(U_i)}{\psi(U_i)} \mid X_j \right].$$

The three terms η_{1j}, η_{2j}, and η_{3j} are all mean zero and have finite variance. They are generally mutually correlated. When θ_0 is known, the term $\eta_{2j} = 0$ and this term is missing from the asymptotic expansion. The term η_{3j} is due to the estimation of ψ even when θ_0 is known.

We next give the distribution theory for the semiparametric estimator $\widehat{\mu}_{4r}(x)$. Let

$$\Psi_F^* = E\left[\frac{\psi'(U_i)}{\psi(U_i)} \{f_0(Z_i, \theta_0) - E[f_0(Z_i, \theta_0)|U_i]\} \gamma_i^* \right] - E\left[\frac{E[f_0(Z_i, \theta_0)|U_i]}{\psi(U_i)} \frac{\partial}{\partial U} E\left[\frac{\partial m}{\partial \theta^\top}(X_i, \theta_0) \mid U_i \right] \right]$$

$$\gamma_i^* = \frac{\partial m}{\partial \theta^\top}(X_i, \theta_0) - E\left[\frac{\partial m}{\partial \theta^\top}(X_i, \theta_0) \mid U_i \right].$$

THEOREM 4. *Suppose that assumptions A1-A4, B1, B2 and C1-C4 hold. Then*

$$\sqrt{n}[\widehat{\mu}_{4r}(x) - \mu_r(x)] \Longrightarrow N(0, \sigma_\eta^{*2}(x)),$$

where $0 < \sigma_\eta^{*2}(x) = \text{var}(\eta_j^*) < \infty$, *with:* $\eta_j^* = \eta_{1j}^* + \eta_{2j}^* + \eta_{3j}^*$, *where* $\eta_{1j}^* = \eta_{1j}$, *while*

$$\eta_{2j}^* = (\Gamma_F - \Psi_F^*)\, \varsigma(Z_j; \theta_0)$$

$$\eta_{3j}^* = -\left(E[f_0(Z_i, \theta_0)|U_i] - E\left[E[f_0(Z_i, \theta_0)|U_i] \right] \right).$$

The three terms η_{1j}^*, η_{2j}^*, and η_{3j}^* are all mean zero and have finite variance. They are generally correlated. When θ_0 is known, the term $\eta_{2j}^* = 0$ and this term is missing from the asymptotic expansion. The term η_{3j}^* is due to the estimation of ψ.

4 Standard Errors and Inference

In the semiparametric case consistent standard errors can be constructed by substituting population quantities by estimated ones along the lines discussed in Newey and McFadden (1994) for finite dimensional parameters. In the nonparametric case, similar methods can be applied (see Härdle and Linton (1994)) for estimating the asymptotic variance; however, in that case the bias is hard to estimate. An alternative approach to inference here is based on the bootstrap. In our case a standard i.i.d. resample from the data set can be shown to work for the nonparametric and semiparametric cases even under our discrete/asymptotically continuous design at least as far as approximating the asymptotic variance, see Shao and Tu (1995), Horowitz (2001), and Mammen (1992). We have taken this approach to inference below due to its simplicity.

5 Efficiency Comparison and Robustness

We have shown that generally $\widehat{\mu}_{2r}(x)$ has smaller mean squared error than $\widehat{\mu}_{1r}(x)$. However, there are other comparisons between the estimators that are also relevant. For example, the estimator $\widehat{\mu}_{1r}(x)$ requires prior knowledge of $h(v \mid x)$, and entails more smoothness than $\widehat{\mu}_{2r}(x)$, as can be seen from the bias expressions given above. On the other hand $\widehat{\mu}_{1r}(x)$ also uses a lower dimensional smoothing operation than $\widehat{\mu}_{2r}(x)$, which may be important in small samples.[12] An advantage of the estimator $\widehat{\mu}_{1r}(x)$ is that it takes the form of a standard nonparametric regression estimator, so known regression bandwidth selection methods can be automatically applied, whereas a comparable theory relevant for $\widehat{\mu}_{2r}(x)$ is not so well developed.

Regarding the semiparametric estimators, it is not possible to provide an efficiency ranking of the two estimators $\widehat{\mu}_{3r}(x)$ and $\widehat{\mu}_{4r}(x)$ uniformly throughout the 'parameter space'. This result partly depends on the choice of $\widehat{\theta}$. It may be possible to develop an efficiency bound for estimation of the function $\mu_r(.)$ by following the calculations of Bickel, Klaassen, Ritov and Wellner (1993, Chapter 5). Since there are no additional restrictions on μ_r, the plug-in estimator with efficient $\widehat{\theta}$ should be efficient. See, e.g., Brown and Newey (1998)

[12] The evidence on the finite sample performance of marginal integration estimators is mixed, see Sperlich, Linton, and Härdle (1999).

6 Extensions

6.1 A Special Case

Suppose that the semiparametric specification A4 holds and that

$$r(w,x) = [\Lambda^{-1}(w)]^k. \tag{15}$$

When Λ is the identity function, this would include many common choices of r in applications. Let $s_k(X,V,Y)$ denote $s_r(X,V,Y)$ with $r(w,x) = [\Lambda^{-1}(w)]^k$. For any k we then have

$$E[(\Lambda^{-1}(W))^k \mid X = x] = E\left[(m(X,\theta_0) - \varepsilon)^k \mid X = x\right] = \sum_{\ell=0}^{k} m(x,\theta_0)^\ell (-1)^{k-\ell} \binom{k}{\ell} E(\varepsilon^{k-\ell})$$

by the binomial expansion. Therefore,

$$E[[\Lambda^{-1}(W)]^k \mid X = x] = \sum_{\ell=0}^{k} m(x,\theta_0)^\ell \alpha_{k\ell},$$

where $\alpha_{k\ell}$, $\ell = 0,\ldots,k$ are unknown parameters depending on the moments of the error distribution. It also follows from Theorem 1 that

$$\lim_{n\to\infty} E\left[s_k(X,V,Y) \mid X = x\right] = \lim_{n\to\infty} E(\Lambda^{-1}(W)^k \mid X = x).$$

We may estimate the nuisance parameters $\alpha_{k\ell}$ along by solving the least squares problem

$$(\widehat{\alpha}_{k0},\ldots,\widehat{\alpha}_{kk}) = \arg\min_{\alpha_{k0},\ldots,\alpha_{kk}} \frac{1}{n}\sum_{i=1}^{n}\left(s_k(X_i,V_i,Y_i) - \sum_{\ell=0}^{k} m(X_i,\widehat{\theta})^\ell \alpha_{k\ell}\right)^2,$$

where $\widehat{\theta}$ is any root-n consistent estimator such as defined in (7). Then let

$$\widehat{\mu}_{6w^k}(x) = \sum_{\ell=0}^{k} m(x,\widehat{\theta})^\ell \widehat{\alpha}_{k\ell}. \tag{16}$$

to estimate $\mu_{w^k}(x)$. Given $\widehat{\theta}$ one has explicit formula for $\widehat{\mu}_{5w^k}(x)$. One could also choose the parameters simultaneously

$$\widehat{\tau} = (\widehat{\theta},\widehat{\alpha}_{k0},\ldots,\widehat{\alpha}_{kk}) = \arg\min_{\alpha_{k0},\ldots,\alpha_{kk},\theta} \frac{1}{n}\sum_{i=1}^{n}\left(s_k(X_i,V_i,Y_i) - \sum_{\ell=0}^{k} m(X_i,\theta)^\ell \alpha_{k\ell}\right)^2,$$

which may or may not be quite convenient.

The asymptotic properties follow from standard theory for nonlinear least squares estimation of nonlinear regression with parameters $\tau = (\theta, \alpha_{k0}, \ldots, \alpha_{kk})$ in the presence of heteroskedasticity [note that $\text{var}[s_{w^k}(X_i, V_i, Y_i)|X_i = x]$ varies with x]. Let $g(\tau) = \sum_{\ell=0}^{k} m(x,\theta)^\ell \alpha_{k\ell}$, $G(\tau) = \partial g(\tau)/\partial\tau$, and

$$\begin{aligned} \Gamma &= E\left[M(X_i)M^\top(X_i)\right] \quad ; \quad \Omega = E\left[M(X_i)M^\top(X_i)\text{var}\left(s_{w^k}(X_i,V_i,Y_i)|X_i\right)\right], \\ \text{where } M(X_i) &= \frac{\partial}{\partial\tau}\left(\sum_{\ell=0}^{k} m(X_i,\theta)^\ell \alpha_{k\ell}\right)_{\tau=\tau_0}. \end{aligned}$$

For identification we require that the matrix Γ be of full rank. It follows that

$$\sqrt{n}(\widehat{\tau} - \tau_0) \Longrightarrow N(\Gamma^{-1}\Omega\Gamma^{-1}),$$

$$\sqrt{n}(\widehat{\mu}_{5w^k}(x) - \mu_{5w^k}(x)) \Longrightarrow N(G^\top\Gamma^{-1}\Omega\Gamma^{-1}G),$$

see for example Pakes and Pollard (1989).

This estimator should work well when k is small, but otherwise a large number of auxiliary parameters $\alpha_{k\ell}$ have to be estimated and this may result in the estimate of $\mu_r(x)$ having a large variance. It is also sensitive to the existence of higher moments.

This method could also be extended to more general class of Λ functions. Suppose that $\Lambda(t) = \sum_{j=0}^{\infty}\psi_j t^j$ for some known coefficients $\{\psi_j\}_{j=1}^{\infty}$. This is true for a large class of Λ functions of interest like the exponential and logarithm. Then $E(W|X=x) = \sum_{j=0}^{\infty}\psi_j\left(m(x,\theta_0)-\varepsilon\right)^j = \sum_{j=0}^{\infty}\widetilde{\psi}_j m(x,\theta_0)^j$ for some coefficients $\widetilde{\psi}_j$ depending on the error moments. In practice, we approximate $E(W|X=x)$ by $\widehat{\mu}(x) = \sum_{j=0}^{L}\widehat{\widetilde{\psi}}_j m(x,\widehat{\theta})^j$, where $L = L(n)$ is some truncation parameter, and where $\widehat{\widetilde{\psi}}_j, \widehat{\theta}$ are estimates obtained from least squares regressions. Other moments can similarly estimated by truncated power series approximations.

6.2 Quantiles

Let $w_q(x)$ denote the q'th conditional quantile of W given $X = x$. It follows immediately from Assumption A.1, in particular equation (1), that

$$w_q(x) = G^{-1}(1-q \mid x), \tag{17}$$

where $G(v \mid x) = E(Y \mid V = v, X = x)$, so we may invert a nonparametric estimator of this expectation to obtain an estimate of $w_q(x)$, for any q such that $1-q \in \text{supp}(V)$, and so will be identified for all quantiles given Assumption A.3. The rate of convergence of $\widehat{w}_q(x) = \widehat{G}^{-1}(1-q \mid x)$ will be slow, because of the high dimension of $\widehat{G}$ and because it behaves like a regression function not a conditional c.d.f.

For semiparametric quantile estimation, if Assumptions A.1 and A.4 hold then

$$q = \Pr[\Lambda[m(X,\theta_0) - \varepsilon] \leq w_q(X) \mid X = x)] = 1 - F_\varepsilon[m(X,\theta_0) - \Lambda^{-1}(w_q(x))], \text{ so}$$

$$w_q(x) = \Lambda[m(X,\theta_0) - F_\varepsilon^{-1}(1-q)]$$

and from Corollary 1, F_ε is obtained by $F_\varepsilon(u) = E(Y \mid U = u)$. Therefore, let $\widehat{U}_i = m(X_i,\widehat{\theta}) - \Lambda^{-1}(V_i)$ and estimate the conditional quantile $w_q(x)$ by

$$\widehat{F_\varepsilon}(u) = \widehat{E}(Y \mid \widehat{U} = u) \quad ; \quad \widetilde{w}_q(x) = \Lambda[m(x,\widehat{\theta}) - \widehat{F_\varepsilon}^{-1}(1-q)],$$

where the function $\widehat{F_\varepsilon}$ is obtained by nonparametrically regressing Y on $\widehat{U}$, and is then numerically inverted to get $\widehat{F_\varepsilon}^{-1}$. This estimator $\widetilde{w}_q(x)$ will converge at a faster rate than the nonparametric quantile estimator $\widehat{w}_q(x)$, because estimation of the quantiles $w_q(x)$ given θ only requires estimation of the one dimensional regression $F_\varepsilon(u) = E(Y \mid U = u)$, instead of the high dimensional $G(v \mid x)$.

The distribution theory for our quantile estimators is immediate. The estimator $\widehat{w}_q(x) = \widehat{G}^{-1}(1-q \mid x)$ has the standard distribution theory for conditional quantile estimation. See, e.g., Chaudhuri (1991). The distribution theory for $\widetilde{w}_q(x) = m(x,\widehat{\theta}) - \widehat{F_\varepsilon}^{-1}(1-q)$ is the same as the distribution theory for $\widetilde{w}_q(x) = m(x,\theta_0) - \widetilde{F_\varepsilon}^{-1}(1-q)$, where

$$\widetilde{F_\varepsilon}(u) = \widehat{E}(Y \mid U = u),$$

which is again a standard one-dimensional conditional quantile estimator. This is because $\widehat{\theta}$ converges at rate root-n, so the estimation error in $\widehat{\theta}$ is asymptotically irrelevant given the slower convergence rate of quantiles.

The quantile estimators converge more slowly than the mean estimators because there is one more dimension in the smoothing or at least one lest degree of averaging. This might be specific to the estimation strategy adopted here, but it seems to be difficult to avoid. For example, one might be tempted to write

$$\text{med}(W|X = x) = \arg\min_\theta E\left[r_\theta(W)|X = x\right],$$

where $r_\theta(w) = |w - \theta|$ and solve the resulting optimization problem to deliver an estimator of $\text{med}(W|X = x)$. However, $r_\theta(w)$ is not differentiable in θ for all w, and even if one uses the a.e. derivative, $\text{sign}(w - \theta)$, the resulting criterion function is not regular enough so that the in the empirical problem one does not obtain asymptotic normality at the same rate as for differentiable r.

7 Numerical Results

7.1 Monte Carlo

We report the results of a small simulation experiment based on a design of Crooker and Herriges (2004). Let

$$W_i = \beta_1 + \beta_2 X_i + \sigma\varepsilon_i,$$

where X_i is uniformly distributed on $[-30, 30]$ and ε_i is standard normal. We take $\beta_1 = 100$ and $\beta_2 = 2$, which guarantees that the mean WTP is equal to 100. We vary the value of $\sigma \in \{5, 10, 25, 50\}$ and sample size $n \in \{100, 300, 500\}$. For our first set of experiments the bid values are five points in $[25, 175]$ if $n = 100$, ten points if $n = 300$, and 15 points if $n = 500$; these points are randomly assigned to individuals i before drawing the other data and so are fixed in repeated experiments. We take $\kappa = 100$. This design was chosen because it permits direct comparison with the parametric and SNP estimators of WTP considered by Crooker and Herriges (2004), at least when $n = 100$ (they did not increase the number of bids with sample size).

In this case $G(v|x) = 1 - \Phi((v - \beta_1 - \beta_2 x)/\sigma)$ and $g(v|x) = \phi((v - \beta_1 - \beta_2 x)/\sigma)/\sigma$, where Φ, ϕ denote the standard normal c.d.f. and density functions respectively. We estimate the moments: $E[W \mid X = x]$, i.e., $r(w, x) = w$, and $\text{std}(W \mid X = x) = \sqrt{E[W^2 \mid X = x] - E^2[W \mid X = x]}$, which corresponds to taking $r(w, x) = (w^2, w)$ and then computing the square root of $r_{w^2} - r_w^2$. Then: $\mu_w(x) = \beta_1 + \beta_2 x$, $\mu_{w^2}(x) = (\beta_1 + \beta_2 x)^2 + \sigma^2$, and $\text{std}(W \mid X = x) = \sigma$.

We compute estimators $\widehat{\mu}_\lambda(.)$ for $\lambda = 1, 2, 3, 4, 6$. In the computation of $\widehat{\mu}_1(.)$ and $\widehat{\mu}_4(.)$ we used a Gaussian kernel and Silverman's rule of thumb bandwidth. This kernel and bandwidth is not likely to be optimal for this problem, but they are convenient and hence fairly widely used choices in practice.

In this design, the estimator $\widehat{\mu}_1(x)$ is prdicted to be approximately unbiased while the predicted bias of $\widehat{\mu}_2(x)$ is small but non-zero.

In Table 1 and 2 we report four different performance measures: root pointwise mean squared error (RPMSE), pointwise mean absolute error (PMAE), root integrated mean squared error (RIMSE), and integrated mean absolute error (IMAE). Crooker and Herriges (2004) only report pointwise results. Like Crooker and Herriges, our pointwise results are calculated at the central point $x = 0$. Thus, their Table 2a ($n = 100$) and Appendix Table 1a ($n = 300$) are directly comparable with a subset of our results. Our conclusions are:

(A1) The performance of our estimators improves as σ decreases and as sample size increases according to all measures: the pointwise measures improve at approximately our theoretical asymptotic rate, while the integrated measures improve much more slowly; the semi-parametric estimators improve more rapidly with sample size.

(A2) For the larger samples, estimator $\widehat{\mu}_4$ performs best according to nearly all measures although for large σ, $\widehat{\mu}_4$ performs almost identically. For smaller sample sizes the ranking is a bit more variable: only $\widehat{\mu}_3$ is never ranked first.

(A4) Our best estimators always perform better than the Crooker and Herriges SNP estimator.

(A5) The estimates of std$(W \mid X = x)$ are subject to much more variability and bias than the estimates of $E[W \mid X = x]$, particularly in the large σ case.

While our estimators seem to work reasonably well in this discrete bid case, we would expect to obtain better results when the bid distribution is actually continuous. We repeated the above experiments with bid distribution uniform on $[25, 175]$ and report the results in Tables 3 and 4. Our conclusions are:

(B1) The performance in the continuous design is somewhat better than in the discrete design. For some designs the pointwise results in Table 1 are better, but the integrated results are always better in Table 3. Note that for the pointwise results the chosen point of evaluation $x = 0$ corresponds to $E[W \mid X = 0] = 100$ and in Table 1 there is a point mass in the distribution of the bids at this point.

(B2) The results for standard deviation estimation are in most cases better in Table 4 than in Table 2.

(B3) The ranking of the estimators is the same in Table 3 as Table 1. Once again $\widehat{\mu}_4$ performs the best in large samples.

7.2 Application

We examine a dataset used in An (2000), which is from a contingent valuation study conducted by Hanemann et al. (1991) to elicit the WTP for protecting wetland habitats and wildlife in California's San Joaquin Valley. Each respondent was assigned a bid value. They were then also given a second bid that was either higher or lower than the first, depending on their acceptance or rejection of the first bid. The total number of bid values in this unfolding bracket design is 14: {25, 30, 40, 55, 65, 75, 80, 110, 125, 140, 170, 210, 250, 375}. The dataset consists of bid responses and some personal characteristics of the respondents. The covariates X are age and number of years resident in California, education and income bracket, and binary indicators of sex, race, and membership in an environmental organization. The sample size, after excluding nonrespondents, incomplete responses, etc., is $n = 518$. The marginal distribution of Y across first bids was $\overline{Y}_1 = 0.396$ and across second bids was $\overline{Y}_2 = 0.581$, while $\overline{V}_1 = 132.4$ and $\overline{V}_2 = 153.9$. The second bid was more likely to receive a yes response, which is consistent with the larger mean value of the bid size. The contingency table is

	$Y_2 = 1$	$Y_2 = 0$
$Y_1 = 1$	131	74
$Y_1 = 0$	170	143

This gives a chi-squared statistic of 4.68, which is to be compared with $\chi^2_{0.05}(1) = 3.84$, so we reject the hypothesis of independence across bids, although not strongly.

The individuals for whom either $Y_1 = 0$ and $Y_2 = 1$ or $Y_1 = 1$ and $Y_2 = 0$ reveal a bound on their willingness to pay, because for these individuals we know their WTP lies between $\min\{V_1, V_2\}$ and $\max\{V_1, V_2\}$. By selecting these 244 individuals we obtain that $E(W)$ lies in the interval $[112.1, 187.1]$. This assumes that the first bids themselves do not influence the behaviour in the second round through, e.g., framing or anchoring effects. We provide some empirical evidence below that this assumption may not hold in our data.

We first consider semiparametric specifications for W, in particular:

$$W = X_i^\top \theta - \varepsilon \text{ and } \log(W) = X_i^\top \theta - \varepsilon,$$

so m is linear and Λ is the identity or the exponential function, respectively. With these specifications we estimate the quantity $\mu_w(x) = E(W \mid X = x)$ using our semiparametric estimators $\widehat{\mu}_j(x)$, $j = 3, 4, 6$. To check for possible framing effects, we estimate this conditional mean WTP separately using first bid data and second bid data. Given that first bids were drawn with close to equal probabilites from a discrete distribution of bids, we assumed that the limiting design density $h(V|X)$ is uniform on the interval $[V_{\min}, V_{\max}]$ (which is not a bad approximation).

In Table 5 we report the sample average of the estimates of $E(W \mid X = X_i)$, denoted $\overline{\widehat{\mu}_j}$, $j = 3, 4, 6$, along with bootstrap confidence intervals. The computation of $\widehat{\mu}_j$ is exactly as described in the simulation section.

	Bid 1		Bid 2	
	Linear	Log-Linear	Linear	Log-linear
$\overline{\widehat{\mu}_3}$	110.480 [101.3,126.0]	112.676 [106.4,126.0]	172.838 [154.3,202.3]	356.771 [317.3,631.3]
$\overline{\widehat{\mu}_4}$	105.611 [97.8,118.1]	104.674 [99.7,115.0]	246.059 [196.8,294.5]	715.210 [380.8,1810.4]
$\overline{\widehat{\mu}_6}$	99.653 [92.9,106.1]	99.653 [94.0,106.1]	143.041 [122.1,164.0]	143.041 [121.4,165.3]

Table 5: Estimates of WTP

Table 6 provides parameter estimates along with their 95% bootstrap confidence intervals, and asterisks indicating significant departure from zero at the 5% level.

	Bid 1		Bid 2	
	Linear	Log Linear	Linear	Log Linear
$YEARCA$	0.3823 [−0.118,0.935]	0.0051 [−0.0011,0.01]	0.5382 [−1.24,1.97]	0.0096 [−0.008,0.03]
$FEMALE$	0.5560 [−12.540,11.75]	0.0105 [−0.14,0.15]	28.290 [−10.5,70.8]	0.5033* [0.073,0.98]
$ln(AGE)$	−13.591 [−37.31,8.78]	−0.159 [−0.5,0.12]	−31.714 [−98.9,32.7]	−0.6081 [−1.73,0.33]
$EDUC$	−2.0237 [−4.98,0.72]	−0.0266 [−0.067,10.01]	1.2919 [−10.66,10.10]	0.0563 [−0.05,0.15]
$WHITE$	6.238 [−9.36,26.07]	0.0211 [−0.17,0.21]	60.2098* [3.0,112.0]	0.5206* [0.04,1.08]
$ENVORG$	1.968 [−15.07,16.49]	0.0423 [−0.17,0.20]	34.8597 [−23.9,88.7]	0.0931 [−0.56,0.66]
$ln(INCOME)$	2.378 [−9.70,12.21]	0.0459 [−0.09,0.17]	40.4140* [6.09,68.93]	0.2769 [−0.15,0.56]

Table 6

The estimated mean WTP based on only first bid data agree quite closely and the confidence intervals are quite narrow. Similar results were obtained for the sample median of $\{\widehat{\mu}_j(X_i)\}_{i=1}^n$ and for the estimates at the mean covariate value $\widehat{\mu}_j(\overline{X})$. The results for the second bid data are rather erratic and generally produce higher mean WTP values. This may be an indicator of framing, shadowing, or anchoring effects, in which hearing the first bid and replying to it affects responses to later bids. See, e.g., McFadden (1994), Green et al. (1998) and Hurd et al. (1998). These results may also be due to small sample problems associated with the survey design, in particular, the distribution of second bids differs markedly from the distribution of first bids, including some far larger bid values. An (2000), using a very different modeling methodology, tests and accepts the hypothesis of no framing effects in these data, though he does report some large differences in coefficient estimates based on data using both bids versus just first bid data. Using different estimators and combining both first and second bid data sets, An (2000) reports WTP at the mean ranging from 155 to 227 (plus one outlier estimate of 1341), which may be compared to our estimates of 99 to 113 for first bid data and 143 to 715 using only second bids.

Finally, we conducted a purely nonparametric analysis with each of the four continuous covariates, one at a time. In Figures 2 and 3 we provide the marginal smooths ($\widehat{\mu}_1(X_i)$) themselves along with a pointwise 95% confidence interval. In contrast to the semiparametric model which assumes a linear or loglinear relationship, these figures from the nonparametric estimator show some nonlinear effects.

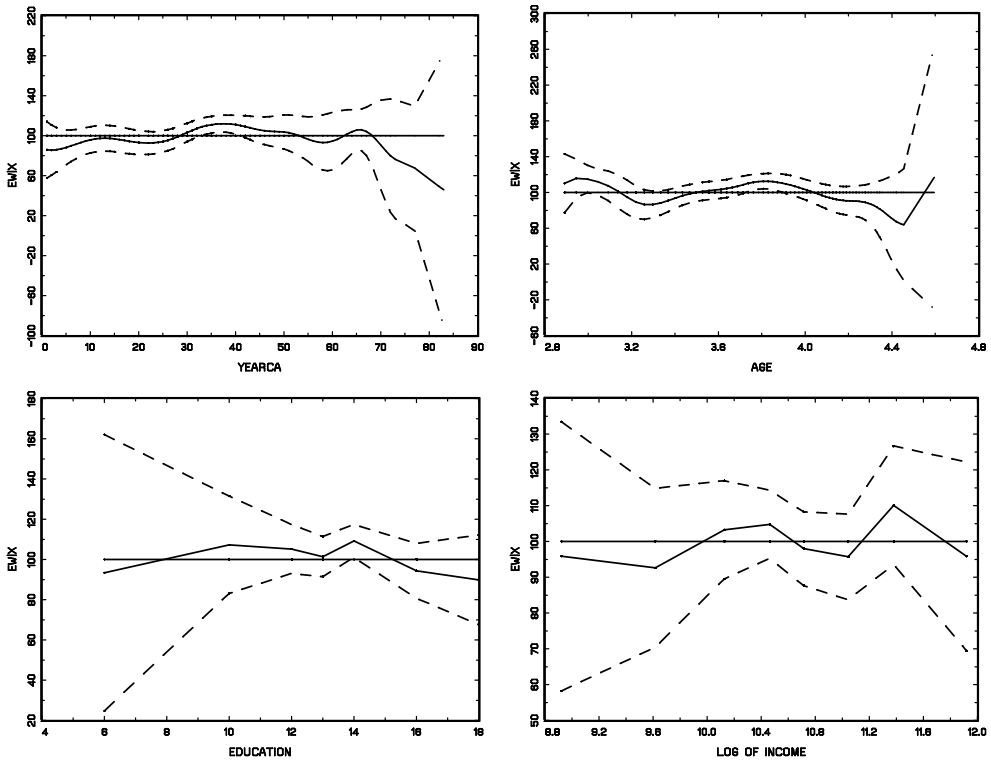

Figure 2. First bid data. Marginal smooths $\widehat{\mu}_1(X_i)$ with pointwise confidence intervals with estimated unconditional mean.

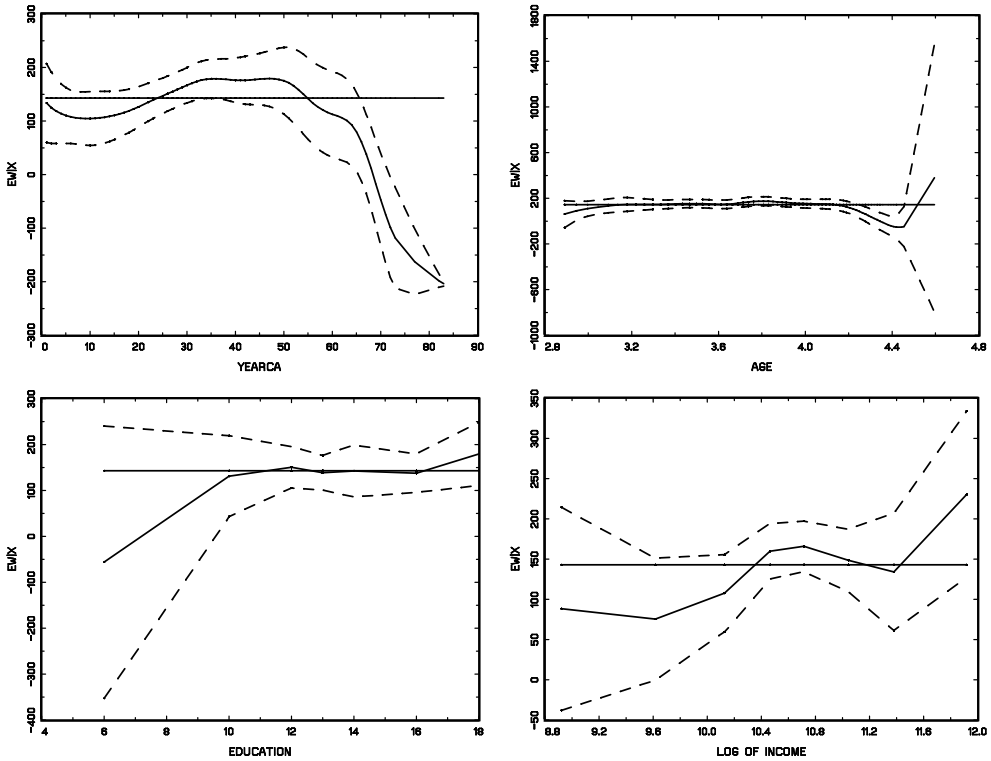

Figure 3. Second bid data. Marginal smooths $\widehat{\mu}_1(X_i)$ with pointwise confidence intervals with estimated unconditional mean.

8 Concluding Remarks

We have provided semiparametric and nonparametric estimators of conditional moments and quantiles of the latent W. The estimators appear to perform well with both simulated and actual data.

We have for convenience assumed throughout that the limiting support of V is bounded. Most of the results here should extend readily to the infinite support case, although some of the estimators may then require asymptotic trimming to deal with issues arising from division by a density estimate when the true density is not bounded away from zero.

The results here show the importance, for both identification and estimation, of experimental designs in which the distribution of bids or test values V possesses at least a fair number of mass points, and ideally is continuous. This should be taken as a recommendation to future designers of contingent valuation experiments. The precision of the estimators also depends in part on the distribution of test values. When designing experiments, one may wish to choose the limiting density h to maximize efficiency based on the variance estimators.

9 Appendix

9.1 Identification With Discrete Bids

The consistency of our estimators shows that moments $\mu_r(x) = E[r(W,X) \mid X = x]$ are nonparametrically identified, given our assumption that as $n \to \infty$, the distribution of V becomes dense in the support of W. As discussed in the introduction, nonparametric identification fails when the limiting support of V is a finite number of mass points, because the conditional distribution of $Y = I(W > V)$ given $X = x, V = v$ only identifies the distribution of $W|X = x$ at each support point v in the support of V, while $E[r(W,X) \mid X = x]$ depends on the distribution of $W|X = x$ at almost every support point w having a nonzero value of $r(w,x)$.

To further motivate our choice of nonparametric identifying assumptions, we show now that if the limiting support of V is a finite number of mass points, then nonparametric identification still fails even given an additive independent error model for W, that is, $W = m(X) - \varepsilon$ with $\varepsilon \perp X$. For simplicity in the proof it is assumed that X is a scalar, m is increasing in X, and V only takes on two values, but the basic logic can be extended to more general cases.

THEOREM 5. *Assume supp(X) is some open or closed interval on the real line, supp(V) =*

$\{-\delta, 0\}$ *for some* $\delta > 0$*, and* $W = m(X) - \varepsilon$ *with* ε *having an unknown, strictly monotonic CDF* $F_\varepsilon(\varepsilon)$ *and* m *strictly monotonically increasing in* X*. Assume* V, X, ε *are mutually independent. Let* $Y = I(W > V)$*. The functions* $m(x)$ *and* $F_\varepsilon(\varepsilon)$ *are not identified given the distribution of* Y *conditional on* V, X*.*

PROOF OF THEOREM 5. Since Y is binary, the distribution of Y given X and V is $G(v \mid x) = E[Y \mid X = x, V = v] = F_\varepsilon[m(x) - v]$. Let $\zeta_0 = \inf[\text{supp}(X)]$, $m_0 = m(\zeta_0)$, and $\zeta_j = m^{-1}(m_0 + j\delta)$ for integers j. Let $\widetilde{m}(x)$ be any strictly monotonic function on $x \in [\zeta_0, \zeta_1]$ such that $\widetilde{m}(\zeta_0) = m_0$ and $\widetilde{m}(\zeta_1) = m_0 + \delta$. Define $\widetilde{F}_\varepsilon(\varepsilon)$ on $\varepsilon \in [m_0, m_0 + \delta]$ by $\widetilde{F}_\varepsilon(\varepsilon) = G[0 \mid \widetilde{m}^{-1}(\varepsilon)]$. Next, define $\widetilde{F}_\varepsilon(\varepsilon)$ on $\varepsilon \in (m_0 + \delta, m_0 + 2\delta]$ by $\widetilde{F}_\varepsilon(\varepsilon) = G[\delta \mid \widetilde{m}^{-1}(\varepsilon - \delta)]$, and define $\widetilde{m}(x)$ on $x \in (\zeta_1, \zeta_2]$ by $\widetilde{m}(x) = \widetilde{F}_\varepsilon^{-1}[G(0 \mid x)]$. Now define $\widetilde{F}_\varepsilon(\varepsilon)$ on $\varepsilon \in (m_0 + 2\delta, m_0 + 3\delta]$ by $\widetilde{F}_\varepsilon(\varepsilon) = G[\delta \mid \widetilde{m}^{-1}(\varepsilon - \delta)]$, and define $\widetilde{m}(x)$ on $x \in (\zeta_2, \zeta_3]$ by $\widetilde{m}(x) = \widetilde{F}_\varepsilon^{-1}[G(0 \mid x)]$. Continue on in this way until the support of x is exhausted By construction, the functions $\widetilde{m}$ and $\widetilde{F}_\varepsilon$ satisfy $G(v \mid x) = \widetilde{F}_\varepsilon[\widetilde{m}(x) - v]$ for all x and v on their support, and hence are observationally equivalent to $m(x)$ and $F_\varepsilon(\varepsilon)$. ■

Notes.

In this theorem, nothing can be identified about the function $m(x)$ (except possibly its endpoints) over the interval $x \in [\zeta_0, \zeta_1]$, since the observable data are consistent with $m(x)$ equalling any regular function over that interval, and the value of $m(x)$ in any other interval is identified only as a function of its unknown values in $[\zeta_0, \zeta_1]$.

The same proof could have been started by letting $\widetilde{F}_\varepsilon(\varepsilon)$ be any regular function with the correct endpoints on $\varepsilon \in [m_0, m_0 + \delta]$, then recovering the corresponding $\widetilde{m}$ on that interval, and proceeding as before. Therefore, the function $\widetilde{F}_\varepsilon$ is also completely unknown (except possibly at endpoints) over an initial interval, and it's values elsewhere are only recoverable as functions of its values in that interval.

The nonidentification here is not just an issue of location or scale. The proof assumes $m(x)$ may be known at two points, $m(\zeta_0)$ and $m(\zeta_1)$, which is equivalent to knowing (or choosing) a location and scale for $m(x)$. Similarly, the proof may be started by assuming $\widetilde{F}_\varepsilon(\varepsilon)$ is known at the two points and $\varepsilon = m_0$ and $\varepsilon = m_0 + \delta$, which is equivalent to knowing (or choosing) a location and scale for $\widetilde{F}$. These functions are therefore not identified up to location and scale.

Here $E[W \mid X = x] = m(x) - E(\varepsilon)$, so the nonidentification of $m(x)$ up to any location shows nonidentification of mean WTP. Other moments are likewise not identified.

This theorem can be applied to show nonidentification of other closely related models. In particular, it implies nonidentification of the nonparametric ordered choice model $Y = jI(\alpha_j < m(x) - \varepsilon \leq a_{j+1})$ for a set of integers j and threshold constants α_j (two of which can be normalized to zero and one to pin down the location of ε and the scaling of both ε and m) It also shows nonidentification

of the model considered by Das (2002), in which $W = m(x) - \varepsilon$ and one only observes which of a few different fixed intervals each observation W lies in. With a partial parameterization, this model is what An (2000) and others call a double bounded dichotomous choice.

It follows from the consistency of our estimator $\widehat{\mu}_{4r}(x)$ (with, e.g., θ estimated using Klein and Spady 1993) that this model can be identified with a fixed discrete design V if $m(x)$ above is parameterized as $m(x, \theta)$ with a known function m and finite parameter vector θ. In this semiparametric specification, continuity of X takes the place of continuity of V.

The implications of Theorem 5 for bid design differ markedly from results on optimal bid design in parametric or semiparametric models. Summarizing Kanninen (1993), Crooker and Herriges (2004) say, in refering to parametric or semiparametric models "estimates of the mean WTP are best with relatively few bid levels."

Some existing estimators implicitly assume identification, such as the sieve estimators proposed by Chen and Randall (1997) and Das (2002), which they apply to data in which v can only take on a finite number of values. Theorem 5 shows that such models are generally not identified.[13]

9.2 Distribution Theory for Nonparametric Estimators

PROOF OF THEOREM 2. The properties of $\widehat{\mu}_{1r}(x)$ are more or less standard, because V_i is part of the variable being smoothed. The only difference is the triangular array nature of the sampling scheme, but given the conditions we made on the way this distribution changes with n, the limiting distribution $H(v|x)$ can replace $H(v|x, n)$ with error of smaller order than the leading term.

We now turn to $\widehat{\mu}_{2r}(x)$. First, we introduce some notation to define the local polynomial estimator $\widehat{G}(v \mid x)$. Following the notation of Masry (1996a,b), let $N_\ell = (\ell + d - 1)!/\ell!(d-1)!$ be the number of distinct d-tuples j with $|j| = \ell$. Arrange these N_ℓ d-tuples as a sequence in a lexicographical order and let ϕ_ℓ^{-1} denote this one-to-one map. Define $\widetilde{X}_i = (V_i, X_i)$ and $\widetilde{x} = (v, x)$, and write $\widehat{G}(v \mid x) = \widehat{G}(\widetilde{x})$ and $G(v \mid x) = G(\widetilde{x})$ for short. We have $\widehat{G}(\widetilde{x}) = e_1^\top M_n^{-1} \Psi_n$, where $e_1 = (1, 0, \ldots, 0)^\top$ is the vector with the one in the first position, $M_n(\widetilde{x})$ and $\Psi_n(\widetilde{x})$ are symmetric $N \times N$ ($N = \sum_{\ell=0}^{p-1} N_\ell \times 1$) matrix and $N \times 1$ dimensional column vector respectively and are defined as

$$M_n(\widetilde{x}) = \begin{bmatrix} M_{n,0,0}(\widetilde{x}) & \cdots & M_{n,0,p-1}(\widetilde{x}) \\ \vdots & \ddots & \vdots \\ M_{n,p-1,0}(\widetilde{x}) & \cdots & M_{n,p-1,p-1}(\widetilde{x}) \end{bmatrix}, \ \Psi_n(\widetilde{x}) = \begin{bmatrix} \Psi_{n,0}(\widetilde{x}) \\ \vdots \\ \Psi_{n,p-1}(\widetilde{x}) \end{bmatrix},$$

[13] Their estimators essentially smooth between the different available test values v to obtain results with uncertain limiting values. Our nonparametric estimators also smooth between test values in an analogous way, but consistency is obtained by having the available bids become dense in the support of W. Crooker and Herriges' (2000) monte carlo design, which we also use, employs this feature of an increasingly fine grid of test values. An (2000) provides a semiparametric model that identifies and estimates the W distribution only at the available bid levels, and explicitly interpolates these estimates to obtain a generally inconsistent estimate of W at the mean.

where $M_{n,|j|,|k|}(\widetilde{x})$ is a $N_{|j|} \times N_{|k|}$ dimensional submatrix with the (l,r) element given by

$$\left[M_{n,|j|,|k|}\right]_{l,r} = \frac{1}{nb^d} \sum_{i=1}^{n} \left(\frac{\widetilde{x} - \widetilde{X}_i}{b} \right)^{\phi_{|j|}(l)+\phi_{|k|}(r)} \widetilde{K}\left(\frac{\widetilde{x} - \widetilde{X}_i}{b} \right),$$

and $\Psi_{n,|j|}(\widetilde{x})$ is a $N_{|j|}$ dimensional subvector whose r-th element is given by

$$\left[\Psi_{n,|j|}\right]_{r} = \frac{1}{nb^d} \sum_{i=1}^{n} \left(\frac{\widetilde{x} - \widetilde{X}_i}{b} \right)^{\phi_{|j|}(r)} \widetilde{K}\left(\frac{\widetilde{x} - \widetilde{X}_i}{b} \right) Y_i.$$

We can write

$$\widehat{G}(\widetilde{x}) - G(\widetilde{x}) = e_1^\top M_n^{-1}(\widetilde{x}) U_n(\widetilde{x}) + e_1^\top M_n^{-1}(\widetilde{x}) B_n(\widetilde{x}). \tag{18}$$

The stochastic term $U_n(\widetilde{x})$ and the bias term $B_n(\widetilde{x})$ are $N \times 1$ vectors

$$U_n(\widetilde{x}) = \begin{bmatrix} U_{n,0}(\widetilde{x}) \\ \vdots \\ U_{n,p-1}(\widetilde{x}) \end{bmatrix}, \; B_n(\widetilde{x}) = \begin{bmatrix} B_{n,0}(\widetilde{x}) \\ \vdots \\ B_{n,d}(\widetilde{x}) \end{bmatrix},$$

where $U_{n,l}(\widetilde{x})$ and $B_{n,l}(\widetilde{x})$ are defined similarly as $\Psi_{n,l}(\widetilde{x})$ so that $U_{n,|j|}(\widetilde{x})$ and $B_{n,|j|}(\widetilde{x})$ are a $N_{|j|}$ dimensional subvectors whose r-th elements are given by:

$$\left[U_{n,|j|}\right]_{r} = \frac{1}{nb^d} \sum_{i=1}^{n} \left(\frac{\widetilde{x} - \widetilde{X}_i}{b} \right)^{\phi_{|j|}(r)} \widetilde{K}\left(\frac{\widetilde{x} - \widetilde{X}_i}{b} \right) \varepsilon_i$$

$$\left[B_{n,|j|}\right]_{r} = \frac{1}{nb^d} \sum_{i=1}^{n} \left(\frac{\widetilde{x} - \widetilde{X}_i}{b} \right)^{\phi_{|j|}(r)} \widetilde{K}\left(\frac{\widetilde{x} - \widetilde{X}_i}{b} \right) \Delta_i(\widetilde{x}),$$

where $\Delta_i(\widetilde{x}) = G(\widetilde{X}_i) - \frac{1}{\mathbf{k}!} \sum_{0 \le |\mathbf{k}| \le p-1} (D^{\mathbf{k}} G)(\widetilde{x})(\widetilde{X}_i - \widetilde{x})^{\mathbf{k}}$, while $\varepsilon_i = Y_i - E(Y_i|\widetilde{X}_i)$ are independent random variables with conditional mean zero and uniformly bounded variances.

The argument is similar to Fan, Mammen, and Härdle (1998, Theorem 1); we just sketch out the extension to our quasi-discrete case. The first part of the argument is to derive a uniform approximation to the denominator in (18). We have

$$\sup_{v \in [\rho_0(x), \rho_1(x)]} |M_n(v,x) - E[M_n(v,x)]| = O_p(a_n), \tag{19}$$

where $a_n = \sqrt{\log n / nb^{d+1}}$. The justification for this comes from Masry (1996a, Theorem 2). Although he assumed continuous density, it is clear from the proofs that the argument goes through in our case. Discreteness of V_i only affects the bias calculation. We calculate $E[M_n(v,x)]$, for simplicity

just the upper diagonal element

$$
\begin{aligned}
& E\widetilde{K}_b(\widetilde{x}-\widetilde{X}_i) \\
= & \int k_b(v-v')K_b\left(x-x'\right)dH(v',x'|n) \\
= & \int k_b(v-v')K_b\left(x-x'\right)dH(v',x') + \int k_b(v-v')K_b\left(x-x'\right)\left[dH(v',x'|n)-dH(v',x')\right].
\end{aligned}
$$

Then using integration by parts for Lebesgue integrals (Carter and van Brunt (2000, Theorem 6.2.2.), for large enough n we have

$$
\int_{\rho_0(x)}^{\rho_1(x)} k_b(v-v')\left[dH(v'|x',n)-dH(v'|x')\right] = -\frac{1}{b^2}\int_{\rho_0(x)}^{\rho_1(x)} k'\left(\frac{v-v'}{b}\right)\left[H(v'|x',n)-H(v'|x')\right]dv',
$$

since the function k is continuous everywhere and the boundary term

$$
\begin{aligned}
\mu_{kH}([\rho_0(x),\rho_1(x)]) = & \; k_b(v-\rho_1(x))\left[H(\rho_1(x)|x',n)-H(\rho_1(x)|x')\right] \\
& -k_b(v-\rho_0(x))\left[H(\rho_0(x)|x',n)-H(\rho_0(x)|x')\right]=0
\end{aligned}
$$

for large enough n, where $\mu_{kH}(A)$ denotes the $H-$measure of the set A. Therefore, by the law of iterated expectation for some constant $C<\infty$,

$$
\begin{aligned}
& \left|\int k_b(v-v')K_b\left(x-x'\right)\left[dH(v',x'|n)-dH(v',x')\right]\right| \\
= & \left|\int k_b(v-v')K_b\left(x-x'\right)\left[dH(v'|x',n)-dH(v'|x')\right]dH(x')\right| \\
= & \left|\frac{1}{b^2}\int k'\left(\frac{v-v'}{b}\right)\left[H(v'|x',n)-H(v'|x')\right]dv'K_b\left(x-x'\right)dH(x')\right| \\
\leq & \sup_{v'}\sup_{|x'-x|\leq b}|H(v'|x',n)-H(v'|x')|\times\frac{1}{b^2}\int|k'\left(\frac{v-v'}{b}\right)|dv'\times\int|K_b\left(x-x'\right)|\,dH(x') \\
\leq & C\left(\frac{1}{b}\sup_{v'}\sup_{|x'-x|\leq b}|H(v'|x',n)-H(v'|x')|\right)\int|k'(t)|dt\int|K\left(u\right)|\,du\times\sup_{|x'-x|\leq b}h(x')=O_p(J^{-1}b^{-1}),
\end{aligned}
$$

by the integrability and smoothness on k. The right hand side does not depend on v so the bound is uniform.

For each j with $0\leq|j|\leq 2(p-1)$, let $\mu_j(\widetilde{K})=\int_{\mathbb{R}^{d+1}}u^j\widetilde{K}(u)du$, $\nu_j(\widetilde{K})=\int_{\mathbb{R}^{d+1}}u^j\widetilde{K}^2(u)du$, and

define the $N \times N$ dimensional matrices M and Γ and $N \times 1$ vector B, where $N = \sum_{\ell=0}^{p-1} N_\ell \times 1$, by

$$M = \begin{bmatrix} M_{0,0} & M_{0,1} & \cdots & M_{0,p-1} \\ M_{1,0} & M_{1,1} & \cdots & M_{1,p-1} \\ \vdots & & & \vdots \\ M_{p-1,0} & M_{p-1,1} & \cdots & M_{p-1,p-1} \end{bmatrix}, \Gamma = \begin{bmatrix} \Gamma_{0,0} & \Gamma_{0,1} & \cdots & \Gamma_{0,p-1} \\ \Gamma_{1,0} & \Gamma_{1,1} & \cdots & \Gamma_{1,p-1} \\ \vdots & & & \vdots \\ \Gamma_{p-1,0} & \Gamma_{p-1,1} & \cdots & \Gamma_{p-1,p-1} \end{bmatrix}, \quad B = \begin{bmatrix} M_{0,p} \\ M_{1,p} \\ \vdots \\ M_{p-1,p} \end{bmatrix},$$

where $M_{i,j}$ and $\Gamma_{i,j}$ are $N_i \times N_j$ dimensional matrices whose (ℓ, m) element are, respectively, $\mu_{\phi_i(\ell)+\phi_j(m)}$ and $\nu_{\phi_i(\ell)+\phi_j(m)}$. Note that the elements of the matrices $M = M(\widetilde{K})$ and $\Gamma = \Gamma(\widetilde{K})$ are simply multivariate moments of the kernel $\widetilde{K}$ and $\widetilde{K}^2$, respectively.

Under the smoothness conditions on $h(v, x)$ we have for all j, k, l, r

$$\frac{1}{b^d} \int \left(\frac{\widetilde{x} - \widetilde{x}'}{b} \right)^{\phi_{|j|}(l)+\phi_{|k|}(r)} \widetilde{K} \left(\frac{\widetilde{x} - \widetilde{x}'}{b} \right) dH(v', x') = h(v, x) \left[M_{|j|,|k|} \right]_{l,r} + O(b)$$

uniformly over v. Therefore,

$$M_n(\widetilde{x}) = h(\widetilde{x})M + O_p(c_n), \tag{20}$$

where $c_n = a_n + b + J^{-1}b^{-1}$, and the error is uniform over v in the support of $H(v|x, n)$. There is an additional term here of order $J^{-1}b^{-1}$ due to the discreteness. This term is of small order under our conditions.

Then $e_1^\top M_n^{-1}(\widetilde{x}) U_n(\widetilde{x}) = e_1^\top M^{-1} U_n(\widetilde{x})/h(\widetilde{x}) + \text{rem}(\widetilde{x})$,where $\text{rem}(\widetilde{x})$ is a remainder term that is $o_p(n^{-1/2}b^{-(d+1)/2})$. By similar arguments we obtain $e_1^\top M_n^{-1}(\widetilde{x}) B_n(\widetilde{x}) = b^p \beta(\widetilde{x}) + \text{rem}(\widetilde{x})$, where $\text{rem}(\widetilde{x})$ is a remainder term that is $o_p(b^p)$ and $\beta(\widetilde{x}) = e_1^\top M^{-1} B G^{(p+1)}(\widetilde{x})$. Therefore, we obtain

$$\widehat{G}(\widetilde{x}) - G(\widetilde{x}) = \frac{1}{h(\widetilde{x})} e_1^\top M^{-1} U_n(\widetilde{x}) + b^p \beta(v, x) + \text{rem}(\widetilde{x}),$$

where $\text{rem}(\widetilde{x})$ is a remainder term that is $o_p(n^{-1/2}b^{-(d+1)/2}) + o_p(b^p)$. We next substitute the leading terms into $\widehat{\mu}_{2r}(x)$, and recall that

$$\widehat{\mu}_{2r}(x) - \mu_r(x) = \int_{\rho_0(x)}^{\rho_1(x)} r'(v, x)[\widehat{G}(v \mid x) - G(v \mid x)]dv.$$

The standard integration argument along the lines of Fan, Mammen, and Härdle (1998) shows that the term $\text{rem}(\widetilde{x})$ can be ignored, and we obtain

$$\widehat{\mu}_{2r}(x) - \mu_r(x) = e_1^\top M^{-1} \overline{U}_n(x) + b^p \overline{\beta}(x) + o_p(n^{-1/2}b^{-d/2}),$$

where $\overline{\beta}(x) = \int \beta(v, x) d\lambda(v)$, while $\overline{U}_n(x)$ is an $N \times 1$ vector

$$\overline{U}_n(x) = \begin{bmatrix} \overline{U}_{n,0}(x) \\ \vdots \\ \overline{U}_{n,p}(x) \end{bmatrix},$$

where $\overline{U}_{n,|j|}(x)$ is an $N_{|j|}$ dimensional subvector whose r-th elements are given by:

$$\left[\overline{U}_{n,|j|}\right]_r = \int u^{\phi^v_{|j|}(r)} k(u)du \frac{1}{nb^d} \sum_{i=1}^{n} \left(\frac{x - X_i}{b}\right)^{\phi^x_{|j|}(r)} K\left(\frac{x - X_i}{b}\right) \frac{r'(V_i, x)}{h(V_i, x)} \varepsilon_i.$$

We can write

$$e_1^\top M^{-1} \overline{U}_n(x) = \frac{1}{nb^d} \sum_{i=1}^{n} L_{d,p}\left(\frac{x - X_i}{b}\right) \frac{r'(V_i, x)}{h(V_i, x)} \varepsilon_i, \text{ where}$$

$$L_{d,p}\left(\frac{x - X_i}{b}\right) = e_1^\top M^{-1} \begin{bmatrix} \vdots \\ \int u^{\phi^v_{|j|}(r)} k(u)du \left(\frac{x - X_i}{b}\right)^{\phi^x_{|j|}(r)} K\left(\frac{x - X_i}{b}\right) \\ \vdots \end{bmatrix}.$$

Under our conditions

$$\begin{aligned}
& E\left[\left(\frac{r'(V_i, x)}{h(V_i, x)}\right)^2 \sigma^2(V_i, X_i) | X_i = x\right] \\
= & \int \left(\frac{r'(v, x)}{h(v, x)}\right)^2 \sigma^2(v, x) dH(v|x, n) \\
= & \int \left(\frac{r'(v, x)}{h(v, x)}\right)^2 \sigma^2(v, x) dH(v|x) + \int \left(\frac{r'(v, x)}{h(v, x)}\right)^2 \sigma^2(v, x) \left[dH(v|x, n) - dH(v|x)\right] \\
= & \int \left(\frac{r'(v, x)}{h(v, x)}\right)^2 \sigma^2(v, x) dH(v|x) + o(1).
\end{aligned}$$

It follows that the asymptotic variance of $\widehat{\mu}_{2r}(x)$ is

$$\begin{aligned}
& \frac{1}{nb^d} E\left[\frac{1}{b^d} L_{d,p}^2\left(\frac{x - X_i}{b}\right) \left(\frac{r'(V_i, x)}{h(V_i, x)}\right)^2 \sigma^2(V_i, X_i)\right] \\
= & \frac{1}{nb^d}\left[\int \frac{1}{b^d} L_{d,p}^2\left(\frac{x - X}{b}\right) \left(\frac{r'(V, x)}{h(V, x)}\right)^2 \sigma^2(V, X) dH(V|X) dH(X) + o(1)\right] \\
\simeq & \frac{1}{nb^d} \left\|L_{d,p}\right\|^2 \int \sigma^2(v, x) \left(\frac{r'(v, x)}{h(v, x)}\right)^2 h(v, x) dv,
\end{aligned}$$

by a change of variables and dominated convergence and taking account of the discreteness error. Furthermore, the central limit theorem holds by the arguments used in Gozalo and Linton (1999, Lemma CLT) and is not affected by the discreteness of V. The quantity $\left\|L_{d,p}\right\|^2$ can also be defined in terms of the basic kernel k.

The properties of

$$\widehat{\mu}_{0r}(x) = -\int_{\rho_0(x)}^{\rho_1(x)} r(v, x) \frac{\partial \widehat{G}(v \mid x)}{\partial v} dv \tag{21}$$

follow similarly. We have

$$\frac{\partial \widehat{G}(v \mid x)}{\partial v} - \frac{\partial G(v \mid x)}{\partial v} = e_v^\top M_{n*}^{-1}(\widetilde{x}) U_{n*}(\widetilde{x}) + e_1^\top M_{n*}^{-1}(\widetilde{x}) B_{n*}(\widetilde{x}),$$

where $e_v^\top = (0, 1, \ldots, 0)$ and $M_{n*}(\widetilde{x})$, $U_{n*}(\widetilde{x})$, and $B_{n*}(\widetilde{x})$ are like $M_n(\widetilde{x})$, $U_n(\widetilde{x})$, and $B_n(\widetilde{x})$ except that they are for one order higher polynomial. Essentially the same integration argument applies to the stochastic part. The bias term arguments are the same except that p is replaced by $p+1$ ■

9.3 Distribution Theory for Semiparametric Quantities

Let E_i denote expectation conditional on Z_i. In the proofs of Theorems 3 and 4 we make use of Lemmas 1 and 2 given below. Define

$$\rho_j(u, \theta) = h[\Lambda(m(X_j, \theta) - u) | X_j] \Lambda'(m(X_j, \theta) - u)$$

and $\psi_\theta(u) = E\rho_j(u, \theta)$ with $\psi(u) = \psi_{\theta_0}(u)$. Then, interchanging differentiation and integration (which is valid under our conditions) we have

$$\psi'(u) = E\frac{\partial \rho_j(u, \theta_0)}{\partial u} = -E\left(\left[h'(\Lambda|X_j)(\Lambda')^2 + h(\Lambda|X_j)\Lambda''\right](m(X_j, \theta_0) - u)\right). \tag{22}$$

Proof of Theorem 3. Recall that

$$\widehat{\mu}_{3r}(x) = r[\Lambda(m(x, \widehat{\theta})), x] + \frac{1}{n}\sum_{i=1}^n \frac{r'[\Lambda(m(x, \widehat{\theta}) - \widehat{U}_i), x]\Lambda'(m(x, \widehat{\theta}) - \widehat{U}_i)[Y_i - 1(\widehat{U}_i > 0)]}{\widehat{\psi}(\widehat{U}_i)},$$

where $\widehat{U}_i = m(X_i, \widehat{\theta}) - \Lambda^{-1}(V_i)$ and

$$\widehat{\psi}(\widehat{U}_i) = \frac{1}{n}\sum_{j=1}^n h[\Lambda(m(X_j, \widehat{\theta}) - \widehat{U}_i)|X_j]\Lambda'(m(X_j, \widehat{\theta}) - \widehat{U}_i)) = \frac{1}{n}\sum_{j=1}^n \rho_j(\widehat{U}_i, \widehat{\theta}).$$

By a geometric series expansion of $1/\widehat{\psi}(\widehat{U}_i)$ about $1/\psi(U_i)$ we can write

$$\widehat{\mu}_{3r}(x) = \frac{1}{n}\sum_{i=1}^n f_1(Z_i, \widehat{\theta}) - \frac{1}{n}\sum_{i=1}^n f_2(Z_i, \theta_0)[\widehat{\psi}(\widehat{U}_i) - \psi(U_i)] \tag{23}$$

$$-\frac{1}{n}\sum_{i=1}^n [f_2(Z_i, \widehat{\theta}) - f_2(Z_i, \theta_0)][\widehat{\psi}(\widehat{U}_i) - \psi(U_i)] \tag{24}$$

$$+\frac{1}{n}\sum_{i=1}^n \frac{r'[\Lambda(m(x, \widehat{\theta}) - \widehat{U}_i), x]\Lambda'(m(x, \widehat{\theta}) - \widehat{U}_i)[Y_i - 1(\widehat{U}_i > 0)]}{\psi^2(U_i)\widehat{\psi}(\widehat{U}_i)}[\widehat{\psi}(\widehat{U}_i) - \psi(U_i)]^2, \tag{25}$$

where

$$f_2(Z_i, \theta) = \frac{r'[\Lambda(m(x, \theta) - U_i(\theta)), x]\Lambda'(m(x, \theta) - U_i(\theta)))[Y_i - 1(U_i(\theta) > 0)]}{\psi^2(U_i)}.$$

The leading terms are derived from (23), while (24) and (25) contain remainder terms.

LEADING TERMS. Lemma 1 implies that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[f_1(Z_i,\widehat{\theta})-Ef_1(Z_i,\theta_0)]=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\Gamma_F\varsigma(Z_i,\theta_0)+[f_1(Z_i,\theta_0)-Ef_1(Z_i,\theta_0)]\}+o_p(1), \quad (26)$$

where $Ef_1(Z_i,\theta_0)=\mu_r(x)$, and $f_1(Z_i,\theta_0)-Ef_1(Z_i,\theta_0)]=f_0(Z_i,\theta_0)-Ef_0(Z_i,\theta_0)]$ due to the cancellation of the common term $r[\Lambda(m(x,\theta_0)),x]$. The stochastic equicontinuity condition of Lemma 1 is verified in a separate appendix, see below.

Let $L(Z_i,Z_j)=\xi_j(U_i)+\Gamma(Z_i)\varsigma(Z_j;\theta_0)$, and

$$\xi_j(u)=\rho_j(u,\theta_0)-E\rho_j(u,\theta_0)$$

$$\Gamma(Z_i)=E_i\left[\zeta_{ij}\frac{\partial m(X_j,\theta_0)}{\partial\theta}\right]-E_i[\zeta_{ij}]\frac{\partial m(X_i,\theta_0)}{\partial\theta},\ \text{where}\ \zeta_{ij}=-\frac{\partial\rho_j(U_i,\theta_0)}{\partial u}.$$

Note that $E_i[\xi_j(U_i)]=0$ but $E_j[\xi_j(U_i)]\neq 0$. We first approximate $n^{-1}\sum_{i=1}^{n}f_2(Z_i,\theta_0)[\widehat{\psi}(\widehat{U}_i)-\psi(U_i)]$ by $n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}f_2(Z_i,\theta_0)L(Z_i,Z_j)$. Specifically, by Lemma 2 and the fact that $E|f_2(Z_i,\theta_0)|<\infty$, we have

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^{n}f_2(Z_i,\theta_0)[\widehat{\psi}(\widehat{U}_i)-\psi(U_i)-\frac{1}{n}\sum_{j=1}^{n}L(Z_i,Z_j)]\right|\\ \leq\ &\frac{1}{n}\sum_{i=1}^{n}|f_2(Z_i,\theta_0)|\times\max_{1\leq i\leq n}\left|[\widehat{\psi}(\widehat{U}_i)-\psi(U_i)-\frac{1}{n}\sum_{j=1}^{n}L(Z_i,Z_j)]\right|\\ =\ &o_p(n^{-1/2}).\end{aligned}$$

Next, letting $\varphi_n(z_1,z_2)=n^{-2}f_2(z_1,\theta_0)L(z_1,z_2)$ we have

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}f_2(Z_i,\theta_0)L(Z_i,Z_j)=\sum_{i=1}^{n}\sum_{j=1}^{n}\varphi_n(Z_i,Z_j),$$

which can be approximated by a second order U-statistic as follows. Letting $p_n(z_1,z_2)=n(n-1)[\varphi_n(z_1,z_2)+\varphi_n(z_2,z_1)]/2$ we have

$$\mathcal{Q}_n=\sum_{i=1}^{n}\sum_{j=1}^{n}\varphi_n(Z_i,Z_j)=\binom{n}{2}^{-1}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}p_n(Z_i,Z_j)+o_p(n^{-1/2}),$$

since $\sum_{i=1}^{n}\varphi_n(Z_i,Z_i)=o_p(n^{-1/2})$. Now p_n is a symmetric kernel, i.e., $p_n(z_1,z_2)=p_n(z_2,z_1)$ and we can apply Lemma 3.1 of Powell, Stock, and Stoker (1989). Letting

$$\widehat{\mathcal{Q}}_n=\frac{2}{n}\sum_{j=1}^{n}\omega_n(Z_j),\ \text{where}\ \omega_n(Z_i)=E_i\left[p_n(Z_i,Z_j)\right],$$

we have $\sqrt{n}(Q_n - \widehat{\mathcal{Q}}_n) = o_p(1)$. It remains to find $\omega_n(Z_i)$. We have

$$2\omega_n(Z_i) = E\left[f_2(Z_j,\theta_0)\Gamma(Z_j)\right]\varsigma(Z_i;\theta_0) + E_i\left[f_2(Z_j,\theta_0)\xi_i(U_j)\right]$$

because $E_i[L(Z_i,Z_j)] = 0$. Furthermore,

$$\begin{aligned} E_j\left[f_2(Z_i,\theta_0)\xi_j(U_i)\right] &= E_j\left[f_2(Z_i,\theta_0)[\rho_j(U_i,\theta_0) - E_i\rho_j(U_i,\theta_0)]\right] \\ &= E_j\left[f_0(Z_i,\theta_0)\frac{[\rho_j(U_i,\theta_0)-\psi(U_i)]}{\psi(U_i)}\right]. \end{aligned}$$

$$\begin{aligned} E\left[f_2(Z_i,\theta_0)\Gamma(Z_i)\right] &= E\left[f_2(Z_i,\theta_0)\left\{E_i\left[\zeta_{ij}\frac{\partial m(X_j,\theta_0)}{\partial\theta}\right] - E_i[\zeta_{ij}]\frac{\partial m(X_i,\theta_0)}{\partial\theta}\right\}\right] \\ &= E\left[\frac{f_0(Z_i,\theta_0)}{\psi(U_i)}\zeta_{ij}\left\{\frac{\partial m(X_j,\theta_0)}{\partial\theta} - \frac{\partial m(X_i,\theta_0)}{\partial\theta}\right\}\right]. \end{aligned}$$

Writing $\zeta_{ij} = E_i\zeta_{ij} + \zeta_{ij} - E_i\zeta_{ij}$, where $E_i\zeta_{ij} = -\psi'(U_i)$, we have

$$\begin{aligned} & E\left[\frac{f_0(Z_i,\theta_0)}{\psi(U_i)}\zeta_{ij}\left\{\frac{\partial m(X_j,\theta_0)}{\partial\theta} - \frac{\partial m(X_i,\theta_0)}{\partial\theta}\right\}\right] \\ = \; & E\left[f_0(Z_i,\theta_0)\frac{\psi'(U_i)}{\psi(U_i)}\left\{\frac{\partial m(X_i,\theta_0)}{\partial\theta} - E\left(\frac{\partial m(X_i,\theta_0)}{\partial\theta}\right)\right\}\right] \\ & + E\left[\frac{f_0(Z_i,\theta_0)}{\psi(U_i)}\left\{\zeta_{ij} - E_i\zeta_{ij}\right\}\left\{\frac{\partial m(X_j,\theta_0)}{\partial\theta} - E\left(\frac{\partial m(X_j,\theta_0)}{\partial\theta}\right)\right\}\right]. \end{aligned}$$

Therefore,

$$\begin{aligned} \widehat{\mathcal{Q}}_n &= E\left[f_0(Z_i,\theta_0)\frac{\psi'(U_i)}{\psi(U_i)}\widetilde{\gamma}_i + \frac{f_0(Z_i,\theta_0)}{\psi(U_i)}\widetilde{\zeta}_{ij}\widetilde{\gamma}_j\right]\frac{1}{n}\sum_{j=1}^n \varsigma(Z_j;\theta_0) \\ &\quad + \frac{1}{n}\sum_{j=1}^n E_j\left[f_0(Z_i,\theta_0)\frac{[\rho_j(U_i,\theta_0)-\psi(U_i)]}{\psi(U_i)}\right]. \end{aligned} \tag{27}$$

We have shown that $n^{-1}\sum_{i=1}^n f_2(Z_i,\theta_0)[\widehat{\psi}(\widehat{U}_i) - \psi(U_i)] = \widehat{\mathcal{Q}}_n + o_p(n^{-1/2})$, where $\widehat{\mathcal{Q}}_n$ is given in (27). This concludes the analysis of the leading terms.

Remainder Terms. By the Cauchy-Schwarz inequality

$$\begin{aligned} & \left|\frac{1}{n}\sum_{i=1}^n [f_2(Z_i,\widehat{\theta}) - f_2(Z_i,\theta_0)][\widehat{\psi}(\widehat{U}_i) - \psi(U_i)]\right| \\ \leq \; & \left(\frac{1}{n}\sum_{i=1}^n [f_2(Z_i,\widehat{\theta}) - f_2(Z_i,\theta_0)]^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^n [\widehat{\psi}(\widehat{U}_i) - \psi(U_i)]^2\right)^{1/2} \\ = \; & O_p(n^{-1}) \end{aligned}$$

from another application of Lemmas 1 and 2.

Therefore,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\frac{r'[\Lambda(m(x,\widehat{\theta})-\widehat{U}_i),x]\Lambda'(m(x,\widehat{\theta})-\widehat{U}_i)[Y_i-1(\widehat{U}_i>0)]}{\psi^2(U_i)\widehat{\psi}(\widehat{U}_i)}[\widehat{\psi}(\widehat{U}_i)-\psi(U_i)]^2\right|$$

$$\leq \frac{\sup_{u\in\mathcal{U}}[\widehat{\psi}(u)-\psi(u)]^2+o_p(n^{-1/2})}{\inf_{u\in\mathcal{U}}\psi^3(u)+o_p(1)}\frac{1}{n}\sum_{i=1}^{n}|r'[\Lambda(m(x,\widehat{\theta})-\widehat{U}_i),x]\Lambda'(m(x,\widehat{\theta})-\widehat{U}_i)|\cdot(|Y_i|+1)$$

$$= o_p(n^{-1/2}).$$

This result used the fact that $\min_{1\leq i\leq n}\widehat{\psi}(\widehat{U}_i)\geq\inf_{u\in\mathcal{U}}\psi(u)+o_p(1)$, which is proved in Lemma 2. Also $\inf_{u\in\mathcal{U}}\psi(u)>0$.

In conclusion, $\sqrt{n}[\widehat{\mu}_{3r}(x)-\mu_r(x;\theta_0)]=n^{-1/2}\sum_{i=1}^{n}\eta_i+o_p(1)$, as required. The asymptotic distribution of $\sqrt{n}[\widehat{\mu}_{3r}(x)-\mu_r(x)]$ follows from the central limit theorem for independent random variables with finite variance.

■

Proof of Theorem 4. By a geometric series expansion we can write

$$\widehat{\mu}^*_{4r}(x;\widehat{\theta}) = \frac{1}{n}\sum_{i=1}^{n}f_1(Z_i,\widehat{\theta})-\frac{1}{n}\sum_{i=1}^{n}f_2(Z_i,\theta_0)[\widetilde{\psi}(\widehat{U}_i)-\psi(U_i)] \quad (28)$$

$$-\frac{1}{n}\sum_{i=1}^{n}[f_2(Z_i,\widehat{\theta})-f_2(Z_i,\theta_0)]\times[\widetilde{\psi}(\widehat{U}_i)-\psi(U_i)] \quad (29)$$

$$+\frac{1}{n}\sum_{i=1}^{n}\frac{r'[\Lambda(m(x,\widehat{\theta})-\widehat{U}_i),x]\Lambda'(m(x,\widehat{\theta})-\widehat{U}_i)[Y_i-1(\widehat{U}_i>0)]}{\psi^2(U_i)\widetilde{\psi}(\widehat{U}_i)}[\widetilde{\psi}(\widehat{U}_i)-\psi(U_i)]^2. \quad (30)$$

The leading terms in this expansion are derived from (28), while (29) and (30) contain remainder terms.

Leading Terms. We make use of Lemma 3 given below. The term $n^{-1}\sum_{i=1}^{n}f_1(Z_i,\widehat{\theta})$ has already been analyzed above. By Lemma 3 we have with probability tending to one for some function d(.) with finite r moments

$$\left|\frac{1}{n}\sum_{i=1}^{n}f_2(Z_i,\theta_0)\left[\widetilde{\psi}(\widehat{U}_i)-\psi(U_i)-\frac{1}{n}\sum_{j=1}^{n}L^*(Z_i,Z_j)\right]\right| \leq \frac{1}{nb^3}\left(\frac{1}{n}\sum_{i=1}^{n}|f_2(Z_i,\theta_0)|d(X_i)\right)$$

$$= O_p(n^{-1}b^{-3}). \quad (31)$$

where $L^*(Z_i,Z_j)=b^{-1}k((U_i-U_j)/b)-\psi(U_i)+\Gamma^*(Z_i)\cdot\varsigma(Z_j,\theta_0)$ and

$$\Gamma^*(Z_i) = \psi(U_i)\left\{\frac{\psi'(U_i)}{\psi(U_i)}\left[\frac{\partial m}{\partial \theta^\top}(X_i,\theta_0) - E\left[\frac{\partial m}{\partial \theta^\top}(X_i,\theta_0) \mid U_i\right]\right] - \overline{m}'_\theta(U_i)\right\}.$$

Under our bandwidth conditions, the right hand side of (31) is $o_p(n^{-1/2})$.

Next,

$$\frac{1}{n}\sum_{i=1}^n f_2(Z_i,\theta_0)\frac{1}{n}\sum_{j=1}^n L^*(Z_i,Z_j) = \sum_{i=1}^n\sum_{j=1}^n \varphi_n(Z_i,Z_j)$$

where

$$\varphi_n(Z_i,Z_j) = \frac{1}{n^2}f_2(Z_i,\theta_0)\left[\frac{1}{b}k\left(\frac{U_i-U_j}{b}\right) - \psi(U_i) + \Gamma^*(Z_i)\cdot\varsigma(Z_j,\theta_0)\right].$$

Note that

$$\begin{aligned} E_i\varphi_n(Z_i,Z_j) &= \frac{1}{n^2}f_2(Z_i,\theta_0)\left[\int \frac{1}{b}k\left(\frac{U_i-u}{b}\right)\psi(u)du - \psi(U_i)\right] \\ &= \frac{1}{n^2}f_2(Z_i,\theta_0)\left[\int k(t)\psi(t+U_ib)dt - \psi(U_i)\right] = O_p(n^{-2}b^2) \end{aligned}$$

uniformly in i. Define $\overline{f}_2(U_i) = E[f_2(Z_i,\theta_0)|U_i]$. Then by iterated expectation

$$n^2E_j\varphi_n(Z_i,Z_j) = E\left[\overline{f}_2(U_i)\frac{1}{b}k\left(\frac{U_i-U_j}{b}\right)\right] - E\left[\overline{f}_2(U_i)\psi(U_i)\right] + E\left[f_2(Z_i,\theta_0)\Gamma^*(Z_i)\right]\cdot\varsigma(Z_j,\theta_0),$$

where, using integration by parts, a change of variable, and dominated convergence,

$$E\left[\overline{f}_2(U_i)\frac{1}{b}k\left(\frac{U_i-U_j}{b}\right)\right] = \int \overline{f}_2(u)\frac{1}{b}k\left(\frac{u-U_j}{b}\right)\psi(u)du = \overline{f}_2(U_j)\psi(U_j) + O_p(b^2)$$

uniformly in i. Note that $\overline{f}_2(U_j)\psi(U_j) = \overline{f}_0(U_j) = E[f_0(Z_j,\theta_0)|U_j]$. Furthermore,

$$\begin{aligned} E\left[f_2(Z_i,\theta_0)\Gamma^*(Z_i)\right] &= E\left[f_0(Z_i,\theta_0)\left\{\frac{\psi'(U_i)\gamma_i^*}{\psi(U_i)} - \frac{\overline{m}'_\theta(U_i)}{\psi(U_i)}\right\}\right] \\ &= E\left[\frac{\psi'(U_i)}{\psi(U_i)}\left\{f_0(Z_i,\theta_0) - \overline{f}_0(U_i)\right\}\gamma_i^*\right] - E\left[\overline{f}_0(U_i)\frac{\overline{m}'_\theta(U_i)}{\psi(U_i)}\right] \end{aligned}$$

by substituting in for f_2 and decomposing $f_0(Z_i,\theta_0) = \overline{f}_0(U_i) + f_0(Z_i,\theta_0) - \overline{f}_0(U_i)$. Using the same U-statististic argument as in the proof of Theorem 3 we obtain

$$\frac{1}{n^2}\sum_{i=1}^n f_2(Z_i,\theta_0)\sum_{j=1}^n L^*(Z_i,Z_j) = \frac{1}{n}\sum_{j=1}^n \omega_n(Z_j) + o_p(n^{-1/2}),$$

where $\omega_n(Z_j) = \overline{f}_0(U_j) - E[\overline{f}_0(U_j)] + E\left[f_2(Z_i,\theta_0)\Gamma^*(Z_i)\right]\varsigma(Z_j)$.

Remainder Terms. First,

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}[f_2(Z_i,\widehat{\theta})-f_2(Z_i,\theta_0)][\widetilde{\psi}(\widehat{U}_i)-\psi(U_i)]\right| \\
\leq\ & \left(\frac{1}{n}\sum_{i=1}^{n}[f_2(Z_i,\widehat{\theta})-f_2(Z_i,\theta_0)]^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}[\widetilde{\psi}(\widehat{U}_i)-\psi(U_i)]^2\right)^{1/2} \\
=\ & o_p(n^{-1/2}).
\end{aligned}$$

Second

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\frac{r'[\Lambda(m(x,\widehat{\theta})-\widehat{U}_i),x]\Lambda'(m(x,\widehat{\theta})-\widehat{U}_i)[Y_i-1(\widehat{U}_i>0)]}{\psi^2(U_i)\widetilde{\psi}(\widehat{U}_i)}[\widetilde{\psi}(\widehat{U}_i)-\psi(U_i)]^2\right| \\
\leq\ & \frac{\sup_{u\in\mathcal{U}}[\widetilde{\psi}(u)-\psi(u)]^2+o_p(n^{-1/2})}{\inf_{u\in\mathcal{U}}\psi^3(u)+o_p(1)}\frac{1}{n}\sum_{i=1}^{n}|r'[\Lambda(m(x,\widehat{\theta})-\widehat{U}_i),x]\Lambda'(m(x,\widehat{\theta})-\widehat{U}_i)|\cdot(|Y_i|+1) \\
=\ & o_p(n^{-1/2}).
\end{aligned}$$

This result used the fact that $\min_{1\leq i\leq n}\widetilde{\psi}(\widehat{U}_i)\geq\inf_{u\in\mathcal{U}}\psi(u)+o_p(1)$, which is proved in Lemma 3. ■

9.4 Subsidiary Results

Define $F_n(\theta)=n^{-1}\sum_{i=1}^{n}f(Z_i,\theta)$ for some function f, and let $F(\theta)=EF_n(\theta)$ and $\Gamma_F=\partial F(\theta_0)/\partial\theta$.

LEMMA 1. *Assume:*

(i) For some vector ς

$$\sqrt{n}(\widehat{\theta}-\theta_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varsigma(Z_i,\theta_0)+o_p(1)$$

where $E[\varsigma(Z_i,\theta_0)]=0$ *and* $\Omega=E[\varsigma(Z_i,\theta_0)\varsigma(Z_i,\theta_0)^\top]<\infty$.

(ii) There exists a finite matrix Γ_F *of full (column) rank such that*

$$\lim_{\|\theta-\theta_0\|\to 0}\frac{\|F(\theta)-\Gamma_F(\theta-\theta_0)\|}{\|\theta-\theta_0\|}=0.$$

(iii) For every sequence of positive numbers $\{\delta_n\}$ *such that* $\delta_n\to 0$,

$$\sup_{\|\theta-\theta_0\|\leq\delta_n}\left\|\sqrt{n}[F_n(\theta)-F(\theta)]-\sqrt{n}[F_n(\theta_0)-F(\theta_0)]\right\|=o_p(1).$$

Then

$$\sqrt{n}[F_n(\widehat{\theta})-F(\theta_0)]\Longrightarrow N(0,V),\ \textit{where}$$

$$V=\mathrm{var}[\Gamma_F\varsigma(Z_i,\theta_0)+f(Z_i,\theta_0)]=\Gamma_F\Omega{\Gamma_F}^\top+\mathrm{var}[f(Z_i,\theta_0)]+2\Gamma_F E\varsigma(Z_i,\theta_0)f(Z_i,\theta_0).$$

See below for a discussion on the verification of (iii).

PROOF. Since $\widehat{\theta}$ is root-n consistent, there exists a sequence $\delta_n \to 0$ such that $\Pr[||\sqrt{n}(\widehat{\theta}-\theta_0)|| > \delta_n] \to 0$ as $n \to \infty$. We can therefore suppose that $||\sqrt{n}(\widehat{\theta}-\theta_0)|| \leq \delta_n$ with probability tending to one. We have

$$\begin{aligned}
\sqrt{n}[F_n(\widehat{\theta}) - F(\theta_0)] &= \sqrt{n}[F(\widehat{\theta}) - F(\theta_0)] + \sqrt{n}[F_n(\widehat{\theta}) - F(\widehat{\theta})] \\
&= \Gamma_F\sqrt{n}(\widehat{\theta}-\theta_0) + \sqrt{n}[F_n(\theta_0) - F(\theta_0)] + o(\|[\sqrt{n}(\widehat{\theta}-\theta_0)\|) \\
&\quad +\sqrt{n}\{[F_n(\widehat{\theta}) - F(\widehat{\theta})] - [F_n(\theta_o) - F(\theta_0)]\} \\
&= \Gamma_F\sqrt{n}(\widehat{\theta}-\theta_0) + \sqrt{n}[F_n(\theta_0) - F(\theta_0)] + o_p(1)[\text{by (ii) and (iii)}] \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\Gamma_F\varsigma(Z_i,\theta_0) + [f(Z_i,\theta_0) - Ef(Z_i,\theta_0)]\} + o_p(1),
\end{aligned}$$

and the result now follows from standard CLT for independent random variables. ■

LEMMA 2. *Suppose that assumptions C1-C3 hold. Then, as* $n \to \infty$

$$\max_{1\leq i\leq n}\left|\widehat{\psi}(\widehat{U}_i) - \psi(U_i) - \frac{1}{n}\sum_{j=1}^{n} L(Z_i,Z_j)\right| = o_p(n^{-1/2}) \tag{32}$$

$$\max_{1\leq i\leq n}\left|\frac{1}{n}\sum_{j=1}^{n} L(Z_i,Z_j)\right| = O_p(n^{-1/2}) \tag{33}$$

$$\min_{1\leq i\leq n}\widehat{\psi}(\widehat{U}_i) \geq \inf_{u\in\mathcal{U}}\psi(u) + o_p(1) \tag{34}$$

where $L(Z_i,Z_j) = \xi_j(U_i) + \Gamma(Z_i)\varsigma(Z_j;\theta_0)$, *and*

$$\Gamma(Z_i) = E_i\left[\zeta_{ij}\frac{\partial m(X_j,\theta_0)}{\partial\theta}\right] - E_i[\zeta_{ij}]\frac{\partial m(X_i,\theta_0)}{\partial\theta}.$$

$$\begin{aligned}
\xi_j(U_i) &= h(\Lambda|X_j)\Lambda'(m(X_j,\theta_0) - U_i) - E_i\left[h(\Lambda|X_j)\Lambda'(m(X_j,\theta_0) - U_i)\right] \\
\zeta_{ij} &= E\left(\left[h'(\Lambda|X_j)(\Lambda')^2 + h(\Lambda|X_j)\Lambda''\right](m(X_j,\theta_0) - U_i)\right).
\end{aligned}$$

PROOF. Regarding (33), the pointwise rate follows by standard central limit theorem for each $Z_i = z$: we have $EL(z,Z_j) = 0$ for each z and $\sup_z \text{var}L(z,Z_j) < \infty$. Then because the function $L(z,Z_j)$ is bounded Lipschitz, the uniformity over z follows from FCLT.

Result (34) follows by an application of the triangle inequality $\min_{1\leq i\leq n}\psi(U_i) \leq \min_{1\leq i\leq n}\widehat{\psi}(\widehat{U}_i) + \max_{1\leq i\leq n}|\widehat{\psi}(\widehat{U}_i) - \psi(U_i)|$, and the fact that $\max_{1\leq i\leq n}|\widehat{\psi}(\widehat{U}_i) - \psi(U_i)| = o_p(1)$ as a consequence of (32) and (33).

Before showing (32) we show that:

$$\max_{1\leq i\leq n}\widehat{U}_i \leq \max_{1\leq i\leq n} U_i + o_p(1) \tag{35}$$

$$\min_{1\leq i\leq n}\widehat{U}_i \geq \min_{1\leq i\leq n} U_i + o_p(1), \tag{36}$$

from which it follows that we can ignore the possibility that $\widehat{U}_i$ lies outside of the support of U_i, i.e., for any event A

$$\begin{aligned}\Pr[\mathcal{A}] &\leq \Pr\left[\mathcal{A} \text{ and } \{\widehat{U}_1,\ldots,\widehat{U}_n\}\subset\mathcal{U}\right] + \Pr\left[\widehat{U}_j\notin\mathcal{U} \text{ for some } j\right] \\ &\leq \Pr\left[\mathcal{A} \text{ and } \{\widehat{U}_1,\ldots,\widehat{U}_n\}\subset\mathcal{U}\right] + o(1) = o(1). \end{aligned}\tag{37}$$

Proof of (35). We have

$$\widehat{U}_i = U_i + \frac{\partial m}{\partial\theta}(X_i,\overline{\theta})(\widehat{\theta}-\theta_0)$$

by the mean value theorem, where $\overline{\theta}$ are intermediate values between $\widehat{\theta}$ and θ_0. Since $\widehat{\theta}$ is root-n consistent, there exists a sequence $\delta_n\to 0$ such that $\Pr[||\widehat{\theta}-\theta_0||\geq\delta_n]\to 0$. Therefore, with probability tending to one

$$|\frac{\partial m}{\partial\theta}(X_i,\overline{\theta})| \leq \sup_{||\theta-\theta_0||\leq\delta_n}|\frac{\partial m}{\partial\theta}(X_i,\theta)| \leq d_1(X_i).$$

Furthermore, applying the Bonferroni and Markov inequalities

$$\Pr\left[\max_{1\leq i\leq n} d_1(X_i) > \epsilon\sqrt{n}\right] \leq n\Pr\left[d_1(X_i)>\epsilon\sqrt{n}\right] \leq n\frac{Ed_1^r(X_i)}{(\epsilon\sqrt{n})^r} = o(1)$$

for any $\epsilon>0$ when $r>2$. This yields (35); (36) follows similarly.

We next prove (32). Define the stochastic process in θ

$$\widehat{\psi}(U_i(\theta)) = \frac{1}{n}\sum_{j=1}^n \rho_j(U_i(\theta),\theta).$$

Then by Taylor expansion

$$\widehat{\psi}(\widehat{U}_i) - \widehat{\psi}(U_i) = \frac{1}{n}\sum_{j=1}^n \frac{\partial\rho_j(U_i(\theta_0),\theta_0)}{\partial\theta^\top}(\widehat{\theta}-\theta_0) + R_{ni}, \tag{38}$$

where the derivative inside the summation is a total derivative defined below, while

$$R_{ni} = \frac{1}{2n}\sum_{j=1}^n (\widehat{\theta}-\theta_0)^\top \frac{\partial^2\rho_j(U_i(\overline{\theta}),\overline{\theta})}{\partial\theta\partial\theta^\top}(\widehat{\theta}-\theta_0),$$

where $\overline{\theta}$ are intermediate values between $\widehat{\theta}$ and θ_0, while:

$$\begin{aligned}
\frac{\partial \rho_j(U_i(\theta),\theta)}{\partial \theta} &= \left[h'(\Lambda|X_j)(\Lambda')^2 + h(\Lambda|X_j)\Lambda''\right](m(X_j,\theta) - U_i(\theta))\left[\frac{\partial m(X_j,\theta)}{\partial \theta} - \frac{\partial m(X_i,\theta)}{\partial \theta}\right] \\
\frac{\partial^2 \rho_j(U_i(\theta),\theta)}{\partial \theta \partial \theta^\top} &= \left[h''(\Lambda|X_j)(\Lambda')^3 + 3h'(\Lambda|X_j)\Lambda'\Lambda'' + h(\Lambda|X_j)\Lambda'''\right](m(X_j,\theta) - U_i(\theta)) \\
&\quad \times \left[\frac{\partial m(X_j,\theta)}{\partial \theta} - \frac{\partial m(X_i,\theta)}{\partial \theta}\right]\left[\frac{\partial m(X_j,\theta)}{\partial \theta} - \frac{\partial m(X_i,\theta)}{\partial \theta}\right]^\top \\
&\quad + \left[h'(\Lambda|X_j)(\Lambda')^2 + h(\Lambda|X_j)\Lambda''\right](m(X_j,\theta) - U_i(\theta))\left[\frac{\partial^2 m(X_j,\theta)}{\partial \theta \partial \theta^\top} - \frac{\partial^2 m(X_i,\theta)}{\partial \theta \partial \theta^\top}\right].
\end{aligned}$$

Applying (37) we have in (38) that with probability tending to one

$$|R_{ni}| \leq ||\widehat{\theta} - \theta_0||^2 \times \frac{1}{n}\sum_{j=1}^n \sup_{||\theta-\theta_0|| \leq \delta_n} \left\| \frac{\partial^2 \rho_j(U_i(\theta),\theta)}{\partial \theta \partial \theta^\top} \right\| \leq O_p(n^{-1}) \times \frac{1}{n}\sum_{j=1}^n D(X_i, X_j)$$

for some measurable function D with finite mean. Therefore, $\max_{1\leq i\leq n} |R_{ni}| = o_p(n^{-1/2})$. We then show that

$$\max_{1\leq i\leq n} \left| \frac{1}{n}\sum_{j=1}^n \frac{\partial \rho_j(U_i(\theta_0),\theta_0)}{\partial \theta^\top} - E_i\left[\frac{\partial \rho_j(U_i(\theta_0),\theta_0)}{\partial \theta^\top}\right] \right| = o_p(1).$$

The pointwise limit follows by the law of large numbers, and the uniformity is obtained by another application of the Bonferroni and Markov inequalities. Therefore, uniformly in i

$$\widehat{\psi}(\widehat{U}_i) - \widehat{\psi}(U_i) = E_i\left[\frac{\partial \rho_j(U_i(\theta_0),\theta_0)}{\partial \theta^\top}\right](\widehat{\theta} - \theta_0) + o_p(n^{-1/2}).$$

We have

$$E_i\left[\frac{\partial \rho_j(U_i(\theta_0),\theta_0)}{\partial \theta^\top}\right] = E_i\left[\zeta_{ij}\frac{\partial m(X_j,\theta_0)}{\partial \theta}\right] - E_i[\zeta_{ij}]\frac{\partial m(X_i,\theta_0)}{\partial \theta}.$$

This is because by the chain rule

$$\begin{aligned}
\frac{\partial \rho_j(U_i(\theta_0),\theta_0)}{\partial \theta} &= \frac{\partial \rho_j(u,\theta)}{\partial \theta}\Big|_{\theta=\theta_0, u=U_i(\theta_0)} + \frac{\partial \rho_j(u,\theta_0)}{\partial u}\Big|_{\theta=\theta_0, u=U_i(\theta_0)} \frac{\partial U_i(\theta)}{\partial \theta}\Big|_{\theta=\theta_0} \\
&= -\frac{\partial \rho_j(u,\theta_0)}{\partial u}\Big|_{\theta=\theta_0, u=U_i(\theta_0)}\left[\frac{\partial m(X_j,\theta_0)}{\partial \theta} - \frac{\partial m(X_i,\theta_0)}{\partial \theta}\right],
\end{aligned}$$

where $\partial \rho_j(u,\theta_0)/\partial u$ was defined in (22).

■

LEMMA 3. *Suppose that assumptions C1-C4 hold. Then with probability tending to one for some*

function d with finite r moments:

$$\max_{1\leq i\leq n}\left|\widetilde{\psi}(\widehat{U}_i)-\psi(U_i)-\frac{1}{n}\sum_{j=1}^{n}L^*(Z_i,Z_j)\right| \leq \frac{k}{nb^3}d(X_i) \tag{39}$$

$$\max_{1\leq i\leq n}\left|\frac{1}{n}\sum_{j=1}^{n}L^*(Z_i,Z_j)\right| = O_p\left\{\left(\frac{\log n}{nb}\right)^{1/2}\right\}+O_p(b^2) \tag{40}$$

$$\min_{1\leq i\leq n}\widetilde{\psi}(\widehat{U}_i) \geq \inf_{u\in\mathcal{U}}\psi(u)+o_p(1) \tag{41}$$

where

$$L^*(Z_i,Z_j)=\frac{1}{b}k\left(\frac{U_i-U_j}{b}\right)-\psi(U_i)+\Gamma^*(Z_i)\cdot\varsigma(Z_j,\theta_0)$$

$$\Gamma^*(Z_i)=\psi'(U_i)\left[\frac{\partial m}{\partial\theta^\top}(X_i,\theta_0)-E\left[\frac{\partial m}{\partial\theta^\top}(X_i,\theta_0)\mid U_i\right]\right]-\psi(U_i)\overline{m}'_\theta(U_i).$$

PROOF. Define

$$\overline{\psi}(U_i)=\frac{1}{nb}\sum_{j=1}^{n}k\left(\frac{U_i-U_j}{b}\right).$$

Making a second order Taylor series expansion we have

$$\widetilde{\psi}(\widehat{U}_i)-\psi(U_i)=T_{ni}+R_{ni}, \tag{42}$$

where

$$T_{ni}=\overline{\psi}(U_i)-\psi(U_i)+\frac{1}{nb^2}\sum_{j=1}^{n}k'\left(\frac{U_i-U_j}{b}\right)\left[\frac{\partial m}{\partial\theta^\top}(X_i,\theta_0)-\frac{\partial m}{\partial\theta^\top}(X_j,\theta_0)\right](\widehat{\theta}-\theta_0)$$

$$\begin{aligned}R_{ni} &= \frac{1}{2nb^3}\sum_{j=1}^{n}k''\left(\frac{U_i^*-U_j^*}{b}\right)\left[\frac{\partial m}{\partial\theta}(X_i,\theta_0)-\frac{\partial m}{\partial\theta}(X_j,\theta_0)\right](\widehat{\theta}-\theta_0)(\widehat{\theta}-\theta_0)^\top\\ &\quad\times\left[\frac{\partial m}{\partial\theta}(X_i,\theta_0)-\frac{\partial m}{\partial\theta}(X_j,\theta_0)\right]^\top\\ &\quad+\frac{1}{nb^2}\sum_{j=1}^{n}k'\left(\frac{U_i-U_j}{b}\right)(\widehat{\theta}-\theta_0)^\top\left[\frac{\partial^2 m}{\partial\theta\partial\theta^\top}(X_i,\theta^*)-\frac{\partial^2 m}{\partial\theta\partial\theta^\top}(X_j,\theta^*)\right](\widehat{\theta}-\theta_0),\end{aligned}$$

where θ^* are intermediate values between $\widehat{\theta}$ and θ_0, and $U_i^*=U_i(\theta^*)$.

We first show that the remainder terms are of smaller order. We have with probability tending to one

$$\begin{aligned}|R_{ni}| &\leq b^{-3}\sup_u|k''(u)|\cdot||\widehat{\theta}-\theta_0||^2\cdot\left(\left\|\frac{\partial m}{\partial\theta}(X_i,\theta_0)\right\|^2+\frac{1}{n}\sum_{j=1}^{n}\left\|\frac{\partial m}{\partial\theta}(X_j,\theta_0)\right\|^2\right)\\ &\quad+b^{-1}||\widehat{\theta}-\theta_0||^2\cdot\frac{1}{nb}\sum_{j=1}^{n}\left|k'\left(\frac{U_i-U_j}{b}\right)\right|(d_1(X_i)+d_2(X_j))\end{aligned}$$

by the Cauchy-Schwarz inequality. Since the function $|k'(u)|$ is Lipschitz continuous, we can apply the uniform convergence results of Masry (1996a):

$$\begin{aligned}
\max_{1\leq i\leq n}\left|\frac{1}{nb}\sum_{j=1}^{n}\left|k'\left(\frac{U_i-U_j}{b}\right)\right| - E_i\left[\left|k'\left(\frac{U_i-U_j}{b}\right)\right|\right]\right| &= O_p\left\{\left(\frac{\log n}{nb}\right)^{1/2}\right\} \\
\max_{1\leq i\leq n}\left|\frac{1}{nb}\sum_{j=1}^{n}\left|k'\left(\frac{U_i-U_j}{b}\right)\right| d_2(X_j) - E_i\left|k'\left(\frac{U_i-U_j}{b}\right)\right| d_2(X_j)\right| &= O_p\left\{\left(\frac{\log n}{nb}\right)^{1/2}\right\},
\end{aligned}$$

since $E[d_2^r(X_j)] < \infty$. Furthermore,

$$\begin{aligned}
E_i\left[\frac{1}{b}\left|k'\left(\frac{U_i-U_j}{b}\right)\right|\right] &= \int |k'(t)|\,\psi(U_i+tb)dt \\
E_i\left[\left|k'\left(\frac{U_i-U_j}{b}\right)\right| d_2(X_j)\right] &= \int |k'(t)|\,\overline{d}_2(U_i+tb)\psi(U_i+tb)dt
\end{aligned}$$

are uniformly bounded, where $\overline{d}_2(U_i) = E[d_2(X_i)|U_i]$. Therefore, for suitable constants and dominating functions

$$|R_{ni}| \leq \frac{k_1}{nb^3}(d_3(X_i)+k_2) + \frac{k_3}{nb}(d_1(X_i)+k_4)$$

with probability tending to one. This gives the result. Furthermore, we have $\max_{1\leq i\leq n} d_l(X_i) = O_p(n^{1/r})$, so that $\max_{1\leq i\leq n}|R_{ni}| = O_p(n^{-1}b^{-3}n^{1/r})$. Provided $n^{(r-2)/r}b^6 \to \infty$, this term is $o_p(n^{-1/2})$. With additional smoothness conditions on k and m, this condition can be substantially weakened.

We now turn to the leading term T_{ni}. By the Masry (1996a) results

$$\max_{1\leq i\leq n}\left|\frac{1}{nb^2}\sum_{j=1}^{n}k'\left(\frac{U_i-U_j}{b}\right)d(X_j) - E_i\left[\frac{1}{b^2}k'\left(\frac{U_i-U_j}{b}\right)\overline{d}(U_j)\right]\right| = O_p\left\{\left(\frac{\log n}{nb^3}\right)^{1/2}\right\}, \tag{43}$$

for any function d with finite moments, where $\overline{d}(U_j) = E[d(X_j)|U_j]$. Under our bandwidth conditions this term is $o_p(1)$. Furthermore, for any twice continuously differentiable function $\overline{d}(u)$ we have

$$\begin{aligned}
&\left|E\left[\frac{1}{b^2}k'\left(\frac{U_i-U_j}{b}\right)\overline{d}(U_j)\mid U_i\right] - [\overline{d}(U_i)\psi(U_i)]'\right| \qquad (44)\\
=\ &\left|\int \frac{1}{b^2}k'\left(\frac{U_i-u}{b}\right)\overline{d}(u)\psi(u)du - [\overline{d}(U_i)\psi(U_i)]'\right| \\
=\ &\left|\int \frac{1}{b}k\left(\frac{U_i-u}{b}\right)[\overline{d}(u)\psi(u)]'du - [\overline{d}(U_i)\psi(U_i)]'\right| \\
=\ &\left|\int k(t)\left([\overline{d}(U_i+tb)\psi(U_i+tb)]' - [\overline{d}(U_i)\psi(U_i)]'\right)dt\right| \\
=\ &O_p(b^2)
\end{aligned}$$

by integration by parts, change of variables and dominated convergence using the symmetry of k. This order is uniform in i by virtue of the boundedness and continuity of the relevant functions. In (43) and (44) take $\overline{d}(u) = 1$ and $\overline{d}(u) = \overline{m}_\theta(u)$, and note that $[\overline{d}(U_i)\psi(U_i)]' = \overline{d}'(U_i)\psi(U_i) + \overline{d}(U_i)\psi'(U_i)$. Therefore,

$$\begin{aligned}
\frac{1}{nb^2}\sum_{j=1}^{n} k'\left(\frac{U_i - U_j}{b}\right)\frac{\partial m}{\partial \theta^\top}(X_i,\theta_0) &= \frac{\partial m}{\partial \theta^\top}(X_i,\theta_0)\psi'(U_i) + o_p(1) \\
\frac{1}{nb^2}\sum_{j=1}^{n} k'\left(\frac{U_i - U_j}{b}\right)\frac{\partial m}{\partial \theta^\top}(X_j,\theta_0) &= \overline{m}'_\theta(U_i)\psi(U_i) + \overline{m}_\theta(U_i)\psi'(U_i) + o_p(1)
\end{aligned}$$

uniformly in i.

In conclusion,

$$\max_{1\le i\le n}|T_{ni} - \frac{1}{n}\sum_{j=1}^{n} L^*(Z_i,Z_j)| = o_p(n^{-1/2}) \ ; \ \max_{1\le i\le n}|R_{ni}| = o_p(n^{-1/2}),$$

which gives the first part of the lemma. Also, we have

$$\max_{1\le i\le n}|\frac{1}{n}\sum_{j=1}^{n} L^*(Z_i,Z_j)| = O_p\left\{\left(\frac{\log n}{nb}\right)^{1/2}\right\} + O_p(b^2),$$

by the Masry results.

The proof of (41) follows as for (34).

■

9.5 Stochastic Equicontinuity Results

We now show that condition (iii) of Lemma 1 is satisfied in our case. Let $\Theta_n(c) = \{\theta\colon \sqrt{n}|\theta - \theta^0| \le c\}$. Since $\sqrt{n}(\widehat{\theta} - \theta^0) = O_p(1)$, for all $\epsilon > 0$ there exists a c_ϵ and an integer n_0 such that for all $n \ge n_0$, $\Pr[\widehat{\theta} \in \Theta_n(c_\epsilon)] \ge 1 - \epsilon$. Define the stochastic process

$$\nu_n(\theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} f(Z_i,\theta) - E[f(Z_i,\theta)], \quad \theta \in \Theta,$$

where

$$f(Z_i,\theta) = r[\Lambda(m(x,\theta)),x] + \frac{r'[\Lambda(m(x,\theta) - U_i(\theta)),x]\Lambda'(m(x,\theta) - U_i(\theta))[Y_i - 1(U_i(\theta) > 0)]}{\psi(U_i)}$$

and define the pseudo-metric $\rho(\theta,\theta') = E([f(Z_i,\theta) - f(Z_i,\theta')]^2)$, on Θ. Under this metric, the parameter space Θ is totally bounded. We are only interested in the behaviour of this process as θ varies in the small set Θ_n. By writing $\theta = \theta^0 + \gamma n^{-1/2}$, we shall make a reparameterization to $\nu_n(\gamma)$, where $\gamma \in \Gamma(c) \subset \mathbb{R}^p$. We establish the following result:

$$\sup_{\gamma\in\Gamma}|\nu_n(\gamma)-\nu_n(0)|=o_p(1) \tag{45}$$

To prove (45) it is sufficient to show a pointwise law of large numbers, e.g., $\nu_n(\gamma)-\nu_n(0)=o_p(1)$ for any $\gamma\in\Gamma$, and stochastic equicontinuity of the process ν_n at $\gamma=0$. The pointwise result is immediate because the random variables are sums of i.i.d. random variables with finite absolute moment and zero mean; the probability limit of $\nu_n(\gamma)$ is the same for all $\gamma\in\Gamma$ by the smoothness of the expected value in γ. To complete the proof of (45) we shall use the following lemma, proved below, which states that ν_n is stochastically equicontinuous in θ. The difficulty in establishing the required equicontinuity arises solely because the function m inside U is nonlinear in θ.

LEMMA SE. *Under the above assumptions, the process $\nu_n(\gamma)$ is stochastically equicontinuous, i.e., for all $\epsilon>0$ and $\eta>0$, there exists $\delta>0$ such that*

$$\limsup_{n\to\infty}\Pr\left[\sup_{\rho(t_1,t_2)<\delta}|\nu_n(t_1)-\nu_n(t_2)|>\eta\right]<\epsilon.$$

PROOF OF LEMMA SE. By a second order Taylor series expansion of $m(Z_i,\theta)$ around $m(Z_i,\theta^0)$:

$$m(Z_i,\theta^0+\gamma n^{-1/2})=m(Z_i,\theta^0)+\frac{1}{\sqrt{n}}\sum_{k=1}^{p}\frac{\partial m}{\partial\theta_k}(Z_i,\theta^0)\gamma_k+\frac{1}{n}\sum_{k=1}^{p}\sum_{r=1}^{p}\frac{\partial^2 m}{\partial\theta_k\partial\theta_r}(Z_i;\overline{\theta})\gamma_k\gamma_r \tag{46}$$

for some intermediate points $\overline{\theta}$. Define the linear approximation to $m(Z_i,\theta^0+\gamma n^{-1/2})$,

$$T(Z_i,\gamma)=m(Z_i,\theta^0)+\sum_{k=1}^{p}\frac{\partial m}{\partial\theta_k}(Z_i,\theta^0)\gamma_k$$

for any γ. By assumption C2, for all k,r, $\sup_{\theta\in\Theta}|\partial^2 m(Z_i,\theta)/\partial\theta_k\partial\theta_r|^2\le d(Z_i)$ with $Ed(Z_i)<\infty$. Therefore, for all $\delta>0$ there exists an $\varepsilon>0$ such that

$$\begin{aligned}\Pr\left[\frac{1}{\sqrt{n}}\max_{i,k,r}\sup_{\theta\in\Theta_n}\left|\frac{\partial^2 m}{\partial\theta_k\partial\theta_r}(Z_i,\theta)\right|>\varepsilon\right] &\le n\sum_{k,r}\Pr\left[\frac{1}{\sqrt{n}}\sup_{\theta\in\Theta_n}\left|\frac{\partial^2 m}{\partial\theta_k\partial\theta_r}(Z_i,\theta)\right|>\varepsilon\right]\\ &\le \frac{\sum_{k,r}E[d(Z_i)]}{\varepsilon^2}\le\delta,\end{aligned}$$

by the Bonferroni and Chebychev inequalities. Therefore, with probability tending to one

$$\max_{1\le i\le n}\left|\frac{1}{n}\sum_{k=1}^{p}\sum_{r=1}^{p}\frac{\partial^2 m}{\partial\theta_k\partial\theta_r}(Z_i;\overline{\theta})\gamma_k\gamma_r\right|\le\frac{\overline{\pi}}{\sqrt{n}}$$

for some $\overline{\pi} < \infty$. Define the stochastic process

$$\nu_{n1}(\gamma, \pi) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \overline{f}(Z_i, \theta_0 + \gamma n^{-1/2}, \pi n^{-1/2}) - E\overline{f}(Z_i, \theta_0 + \gamma n^{-1/2}, \pi n^{-1/2})$$

on $\gamma \in \Gamma$ and $\pi \in \Pi = [0, \overline{\pi}]$, where

$$\begin{aligned}
& \overline{f}(Z_i, \theta_0 + \gamma n^{-1/2}, \pi n^{-1/2}) \\
= \; & r[\Lambda(m(x, \theta_0 + \gamma n^{-1/2})), x] \\
& + \frac{r'[\Lambda(m(x, \theta_0 + \gamma n^{-1/2}) - U_i(\theta_0 + \gamma n^{-1/2})), x] \Lambda'(m(x, \theta_0 + \gamma n^{-1/2}) - U_i(\theta_0 + \gamma n^{-1/2}))}{\psi(U_i)} \\
& \times [Y_i - 1(T(Z_i, \gamma n^{-1/2}) + \frac{\pi}{\sqrt{n}} > 0)]
\end{aligned}$$

It suffices to show that $\nu_{n1}(\gamma, \pi)$ is stochastically equicontinuous in γ, π, and the deterministic centering term is of smaller order. The latter argument is a standard Taylor expansion. The argument for $\nu_{n1}(\gamma, \pi)$ is very similar to that contained in Sherman (1993) because we have a linear index structure in this part. One can apply Lemma 2.12 in Pakes and Pollard (1989). ■

References

[1] An, M.Y. (2000). "A Semiparametric Distribution for Willingness to Pay and Statistical Inference with Dichotomous Choice CV Data,"

[2] Andrews, D.W.K. (1994): "Asymptotics for Semiparametric Econometric Models by Stochastic Equicontinuity." Econometrica 62, 43-72.

[3] Bickel, P.J., C.A.J. Klaassen, J. Ritov, and J. Wellner (1993), *Efficient and Adaptive Estimation for Semiparametric Models*. Springer: Berlin.

[4] Brown, B.W. and W.K. Newey (1998): "Efficient Semiparametric Estimation of Expectations," *Econometrica,* 66, 453-464.

[5] Carter, N., and B. van Brunt (2000): *The Lebesgue Stieltjes Integral*. Springer, Berlin.

[6] Chaudhuri, P. (1991). "Nonparametric estimates of regression quantiles and their local Bahadur representation," *Annals of Statistics* **19**, 760-777.

[7] Chen, H. and A. Randall (1997): "Semi-nonparametric Estimation of Binary Response Models With an Application to Natural Resource Valuation, *Journal of Econometrics*, 76, 323-340.

[8] Coppejans, M. (2003): "Effective Nonparametric Estimation in the Case of Severly Discretized Data," Duke University.

[9] Creel, M., and J. Loomis (1997): "Semi-nonparametric Distribution-free Dichotomous Choice Contingent Valuation," *Journal of Environmental Economics and Management*, 32, 341-358.

[10] Crooker, J.R., and J.A. Herriges (2004): "Parametric and Semi-Nonparametric Estimation of Willingness-to-Pay in the Dichotomous Choice Contingent Valuation Framework," Environmental and Resource Economics 27, 451-480.

[11] Das, M. (2002), "Minimum Distance Estimators for Nonparametric Models With Grouped Dependent Variables," unpublished manuscript.

[12] Delgado, M., and J. Mora (1995): "Nonparametric and Semiparametric Estimation with Discrete Regressors," *Econometrica,* 63, 1477-1484.

[13] Gozalo, P., and O.B. Linton (1999): "Local Nonlinear Least Squares: Using Parametric Information in Nonparametric Regression." *Journal of Econometrics,* 99, 63-106.

[14] Hardle, W., and O.B. Linton (1994), "Applied Nonparametric Methods," *The Handbook of Econometrics*, vol. IV*,* eds. D.F. McFadden and R.F. Engle III. North Holland.

[15] Hengartner, N., and O. Linton (1996): "Nonparametric regression estimation at design poles and zeros," *The Canadian Journal of Statistics* 24, 583-591.

[16] Ho, K. and P.K. Sen (2000): "Robust Procedures For Bioassays and Bioequivalence Studies," *Sankhya, Ser. B,* 62, 119-133.

[17] Kanninen, B. (1993), "Dichotomous Choice Contingent Valuation," *Land Economics*, 69, 138-146.

[18] Klein, R. and R. H. Spady (1993), "An efficient Semiparametric Estimator for Binary Response Models," *Econometrica* 61, 387-421.

[19] Lewbel, A. (1997), "Semiparametric Estimation of Location and Other Discrete Choice Moments," *Econometric Theory,* 13, 32-51.

[20] Lewbel, A. (2000), "Semiparametric Qualitative Response Model Estimation With Unknown Heteroscedasticity or Instrumental Variables," *Journal of Econometrics* 97, 145-177.

[21] Linton, O. and J.P. Nielsen (1995): "A kernel method of estimating structured nonparametric regression based on marginal integration," *Biometrika*, 82, 93-100.

[22] Lu, Z-Q. (2002): "Nonparametric regression with Singular Design," Kowloon.

[23] Manski, C. and E. Tamer, (2000), "Inference on Regressions with Interval Data on a Regressor or Outcome," *Econometrica*, 70, 519-546.

[24] Matzkin, R., (1992), "Non-parametric and Distribution-free Estimation of the Binary Threshold Crossing and the Binary Choice Models," *Econometrica* 60, 239-270.

[25] Masry, E. (1996a), "Multivariate local polynomial regression for time series: Uniform strong consistency and rates," *J. Time Ser. Anal.* 17, 571-599.

[26] Masry, E., (1996b), "Multivariate regression estimation: Local polynomial fitting for time series. *Stochastic Processes and their Applications* 65, 81-101.

[27] McFadden, D. (1994), "Contingent Valuation and Social Choice," *American Journal of Agricultural Economics,* 76, 4.

[28] McFadden, D. (1998), "Measuring Willingness-to-Pay for Transportation Improvements, in T. Gärling, T. Laitila, and K. Westin (eds.),Theoretical Foundations of Travel Choice Modeling, 339-364, Elsevier Science: Amsterdam.

[29] Newey, W. K. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

[30] Newey, W. K. and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in Handbook of Econometrics, vol. iv, ed. by R. F. Engle and D. L. McFadden, pp. 2111-2245, Amsterdam: Elsevier.

[31] Pakes, A. and D. Pollard. (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57, 1027-57.

[32] Racine, J., and Q. Li (2002): "Nonparametric estimation of regression functions with categorical and continuous data," Manuscript, College Station.

[33] Ramgopal, P., P.W. Laud, and A.F.M. Smith. (1993) "Nonparametric Bayesian Bioassay With Prior Constraints on the Shape of the Potency Curve," *Biometrika*, 80, 489-498.

[34] Sherman, R. P. (1993): "The Limiting Distribution of the Maximum Rank Correlation Estimator," *Econometrica*, 61, 123-37.

[35] Shorack, G. R. and J. A. Wellner, (1986): Empirical Processes With Applications To Statistics, John Wiley and sons.

[36] Silverman, B. (1986): *Density estimation for statistics and data analysis*. London, Chapman and Hall.

[37] Sperlich, S, O.B. Linton, and W. Härdle, (1999): "A Simulation comparison between the Backfitting and Integration methods of estimating Separable Nonparametric Models," *TEST,* 8, 419-458.

[38] Stone, C.J. (1982): "Optimal global rates of convergence for nonparametric regression." *Annals of Statistics* 8, 1040-1053.

[39] Tjostheim, D. and B. H. Auestad (1994): "Nonparametric Identification of Nonlinear Time Series: Projections," *Journal of the American Statistical Association,* 89, 1398-1409.

[40] Wang, M-C, and J. Van Ryzin (1981): "A class of smooth estimators for discrete distributions," *Biometrika* 68, 301-309.

TABLES AND FIGURES

		$\sigma = 5$			$\sigma = 10$			$\sigma = 25$			$\sigma = 50$		
		n=100	n=300	n=500	n=100	n=300	n=500	n=100	n=300	n=500	n=100	n=300	n=500
RPMSE	$\widehat{\mu}_1$	4.04	3.86	2.39	4.20	4.54	3.10	6.51	5.87	4.25	10.54	7.11	5.88
	$\widehat{\mu}_2$	4.52	3.66	2.16	4.49	4.30	2.69	6.37	5.36	3.73	9.78	6.37	5.02
	$\widehat{\mu}_3$	5.65	2.20	1.64	5.64	2.69	1.97	8.09	3.86	2.90	10.69	4.60	3.64
	$\widehat{\mu}_4$	4.80	1.77	1.25	4.75	2.15	1.59	6.42	3.38	2.43	8.42	4.53	3.32
	$\widehat{\mu}_6$	4.77	3.89	2.07	5.03	3.69	2.05	6.05	3.89	2.43	8.37	4.32	3.28
PMAE	$\widehat{\mu}_1$	3.20	3.20	1.89	3.32	3.70	2.47	5.23	4.65	3.34	8.45	5.69	4.61
	$\widehat{\mu}_2$	3.59	3.07	1.73	3.60	3.57	2.13	5.11	4.25	3.02	7.86	5.09	3.92
	$\widehat{\mu}_3$	4.50	1.77	1.31	4.54	2.12	1.59	6.59	3.01	2.29	8.65	3.67	2.91
	$\widehat{\mu}_4$	3.75	1.44	0.99	3.76	1.73	1.28	5.15	2.65	1.90	6.74	3.59	2.62
	$\widehat{\mu}_6$	3.82	3.29	1.63	4.06	3.07	1.60	4.85	3.20	1.93	6.98	3.54	2.67
RIMSE	$\widehat{\mu}_1$	14.39	7.89	5.81	14.29	7.88	5.76	14.84	8.58	6.36	17.29	11.72	9.59
	$\widehat{\mu}_2$	12.85	5.29	2.50	13.22	5.60	3.06	14.18	7.21	4.81	16.60	10.90	8.94
	$\widehat{\mu}_3$	12.28	4.76	3.35	12.21	5.16	3.46	13.52	6.39	4.32	16.12	9.83	7.81
	$\widehat{\mu}_4$	11.90	4.58	3.18	11.80	4.90	3.27	12.56	6.12	4.02	14.65	9.79	7.66
	$\widehat{\mu}_6$	11.83	5.72	3.58	11.80	5.75	3.50	12.32	6.41	4.02	14.62	9.67	7.65
IMAE	$\widehat{\mu}_1$	10.88	5.71	4.10	10.77	5.90	4.17	11.08	6.46	4.74	13.41	8.98	7.33
	$\widehat{\mu}_2$	10.54	4.17	1.92	10.76	4.47	2.34	11.10	5.61	3.68	13.05	8.56	6.94
	$\widehat{\mu}_3$	9.44	3.59	2.49	9.52	3.88	2.62	10.61	4.93	3.33	12.71	7.68	6.13
	$\widehat{\mu}_4$	9.12	3.36	2.33	9.16	3.59	2.45	9.77	4.69	3.05	11.53	7.69	5.98
	$\widehat{\mu}_6$	9.32	4.50	2.73	9.39	4.47	2.65	9.72	4.99	3.11	11.74	7.64	6.04

Table 1. Estimation of conditional mean in discrete bid design; 500 replications;

		$\sigma = 5$			$\sigma = 10$			$\sigma = 25$			$\sigma = 50$		
		n=100	n=300	n=500	n=100	n=300	n=500	n=100	n=300	n=500	n=100	n=300	n=500
RPMSE	$\widehat{\mu}_1$	8.95	7.43	7.12	7.40	5.03	5.35	6.43	4.13	3.87	9.89	8.17	7.32
	$\widehat{\mu}_2$	16.39	14.51	12.89	13.00	10.93	9.97	6.93	5.05	5.08	8.73	7.29	6.64
	$\widehat{\mu}_3$	6.75	2.65	2.08	5.43	2.53	2.52	5.54	3.14	3.18	7.68	5.29	5.12
	$\widehat{\mu}_4$	6.60	2.13	1.39	5.15	1.98	1.49	5.47	2.56	2.14	8.27	5.33	4.87
	$\widehat{\mu}_6$	27.26	13.86	15.33	24.38	14.16	13.71	19.55	15.42	13.01	19.92	15.00	10.80
PMAE	$\widehat{\mu}_1$	8.02	7.03	6.95	6.29	4.36	4.91	5.11	3.32	3.06	7.86	6.86	6.34
	$\widehat{\mu}_2$	16.20	14.43	12.85	12.69	10.77	9.86	5.58	4.22	4.43	7.04	6.10	5.69
	$\widehat{\mu}_3$	5.64	2.04	1.60	4.51	1.97	2.08	4.46	2.52	2.66	6.11	4.45	4.60
	$\widehat{\mu}_4$	5.41	1.59	1.05	4.19	1.54	1.18	4.39	2.03	1.66	6.60	4.42	4.09
	$\widehat{\mu}_6$	23.28	10.77	12.42	21.01	12.25	12.08	17.20	12.80	10.69	14.67	10.98	8.19
RIMSE	$\widehat{\mu}_1$	17.56	11.35	8.11	14.23	9.73	7.68	10.07	9.06	8.83	13.02	13.08	13.61
	$\widehat{\mu}_2$	22.07	14.37	9.93	18.01	11.22	7.69	10.16	6.61	5.86	10.82	11.03	11.78
	$\widehat{\mu}_3$	6.75	2.65	2.08	5.43	2.53	2.52	5.54	3.14	3.18	7.68	5.29	5.12
	$\widehat{\mu}_4$	6.60	2.13	1.39	5.15	1.98	1.49	5.47	2.56	2.14	8.27	5.33	4.87
	$\widehat{\mu}_6$	21.78	14.62	11.68	19.31	13.37	11.36	17.32	14.33	13.35	19.93	17.30	17.30
IMAE	$\widehat{\mu}_1$	15.98	9.62	6.99	12.74	8.23	6.55	7.52	6.44	5.83	9.75	10.18	10.96
	$\widehat{\mu}_2$	21.59	13.37	9.19	17.40	10.24	7.03	8.11	5.27	4.30	8.43	8.94	9.92
	$\widehat{\mu}_3$	5.64	2.04	1.60	4.51	1.97	2.08	4.46	2.52	2.66	6.11	4.45	4.60
	$\widehat{\mu}_4$	5.41	1.59	1.05	4.19	1.54	1.18	4.39	2.03	1.66	6.60	4.42	4.09
	$\widehat{\mu}_6$	17.95	11.24	9.09	16.41	11.33	9.84	14.52	11.34	10.27	14.24	12.64	12.72

Table 2. Estimation of conditional standard deviation in discrete bid design; 500 replications;

		$\sigma = 5$			$\sigma = 10$			$\sigma = 25$			$\sigma = 50$		
		n=100	n=300	n=500	n=100	n=300	n=500	n=100	n=300	n=500	n=100	n=300	n=500
RPMSE	$\widehat{\mu}_1$	5.64	3.30	2.58	6.34	3.79	3.07	8.57	5.42	4.46	11.39	7.25	5.90
	$\widehat{\mu}_2$	5.02	2.90	2.32	5.60	3.33	2.68	7.57	4.83	3.94	10.20	6.39	5.24
	$\widehat{\mu}_3$	4.40	2.24	1.68	5.10	2.70	2.08	6.63	3.71	2.81	8.01	4.55	3.44
	$\widehat{\mu}_4$	3.51	1.66	1.23	4.10	2.10	1.62	5.67	3.14	2.36	7.89	4.42	3.34
	$\widehat{\mu}_6$	5.85	3.36	2.59	5.96	3.39	2.65	6.39	3.72	2.88	7.57	4.34	3.31
PMAE	$\widehat{\mu}_1$	4.41	2.61	2.06	5.00	3.01	2.44	6.83	4.31	3.55	9.07	5.74	4.70
	$\widehat{\mu}_2$	4.00	2.31	1.84	4.43	2.66	2.13	6.06	3.84	3.15	8.16	5.12	4.18
	$\widehat{\mu}_3$	3.43	1.75	1.32	4.07	2.14	1.66	5.29	2.96	2.24	6.37	3.63	2.74
	$\widehat{\mu}_4$	2.76	1.32	0.98	3.26	1.68	1.29	4.52	2.51	1.88	6.22	3.51	2.65
	$\widehat{\mu}_6$	4.66	2.67	2.07	4.76	2.71	2.11	5.11	2.97	2.30	6.05	3.45	2.65
RIMSE	$\widehat{\mu}_1$	12.41	7.59	6.02	12.37	7.56	6.06	13.03	7.99	6.53	15.88	11.22	9.75
	$\widehat{\mu}_2$	7.85	4.42	3.42	8.30	4.80	3.77	10.34	6.39	5.19	14.36	10.33	9.16
	$\widehat{\mu}_3$	8.14	4.57	3.51	8.44	4.69	3.70	9.64	5.50	4.35	12.68	9.06	8.07
	$\widehat{\mu}_4$	7.70	4.32	3.32	7.89	4.38	3.47	9.01	5.14	4.07	12.61	9.00	8.03
	$\widehat{\mu}_6$	9.02	5.22	4.03	9.00	5.12	4.05	9.47	5.51	4.39	12.41	8.96	8.01
IMAE	$\widehat{\mu}_1$	8.88	5.44	4.33	8.97	5.48	4.40	9.73	5.97	4.89	12.11	8.56	7.50
	$\widehat{\mu}_2$	6.01	3.40	2.64	6.35	3.69	2.90	7.91	4.90	4.00	11.01	8.01	7.19
	$\widehat{\mu}_3$	6.04	3.39	2.61	6.39	3.56	2.81	7.48	4.27	3.37	9.80	7.08	6.37
	$\widehat{\mu}_4$	5.62	3.14	2.42	5.87	3.26	2.59	6.92	3.95	3.11	9.76	7.02	6.34
	$\widehat{\mu}_6$	6.95	4.03	3.11	6.99	3.97	3.14	7.37	4.27	3.41	9.61	6.98	6.33

Table 3. Estimation of conditional mean in continuous design; 10,000 replications;

		$\sigma = 5$			$\sigma = 10$			$\sigma = 25$			$\sigma = 50$		
		n=100	n=300	n=500	n=100	n=300	n=500	n=100	n=300	n=500	n=100	n=300	n=500
RPMSE	$\widehat{\mu}_1$	9.65	7.43	6.46	7.51	5.33	4.55	6.74	4.42	3.62	10.70	7.80	7.18
	$\widehat{\mu}_2$	16.71	14.02	12.48	13.05	10.66	9.37	6.92	5.30	4.63	9.60	7.08	6.62
	$\widehat{\mu}_3$	5.20	2.90	2.12	5.20	2.81	2.36	5.44	3.47	3.00	7.95	5.62	5.09
	$\widehat{\mu}_4$	4.31	2.16	1.46	4.14	1.95	1.42	4.98	2.75	2.08	8.67	5.71	4.89
	$\widehat{\mu}_6$	20.62	15.39	13.24	19.29	14.55	12.71	20.71	15.83	13.98	24.09	15.72	12.25
PMAE	$\widehat{\mu}_1$	8.90	7.10	6.24	6.42	4.71	4.05	5.36	3.55	2.87	8.69	6.51	6.17
	$\widehat{\mu}_2$	16.52	13.94	12.43	12.72	10.50	9.25	5.56	4.43	3.94	7.92	5.93	5.67
	$\widehat{\mu}_3$	4.02	2.19	1.64	4.03	2.24	1.94	4.26	2.79	2.47	6.51	4.78	4.51
	$\widehat{\mu}_4$	3.32	1.59	1.10	3.19	1.54	1.12	3.93	2.19	1.67	7.08	4.71	4.11
	$\widehat{\mu}_6$	15.34	11.92	10.46	16.04	12.58	11.20	18.36	13.20	11.29	18.27	11.34	9.23
RIMSE	$\widehat{\mu}_1$	12.18	9.98	9.02	11.03	8.99	8.09	12.29	9.37	8.26	18.70	14.29	12.92
	$\widehat{\mu}_2$	14.69	12.28	11.12	12.07	9.69	8.58	9.88	6.81	5.64	16.49	12.26	11.16
	$\widehat{\mu}_3$	5.20	2.90	2.12	5.20	2.81	2.36	5.44	3.47	3.00	7.95	5.62	5.09
	$\widehat{\mu}_4$	4.31	2.16	1.46	4.14	1.95	1.42	4.98	2.75	2.08	8.67	5.71	4.89
	$\widehat{\mu}_6$	18.69	13.95	12.18	17.50	13.44	11.73	19.04	14.56	12.87	24.26	18.15	15.80
IMAE	$\widehat{\mu}_1$	10.07	8.42	7.63	9.38	7.59	6.77	9.28	6.46	5.45	14.00	10.99	10.34
	$\widehat{\mu}_2$	13.48	11.51	10.49	11.16	9.05	8.02	7.38	4.97	4.15	12.55	9.82	9.34
	$\widehat{\mu}_3$	4.02	2.19	1.64	4.03	2.24	1.94	4.26	2.79	2.47	6.51	4.78	4.51
	$\widehat{\mu}_4$	3.32	1.59	1.10	3.19	1.54	1.12	3.93	2.19	1.67	7.08	4.71	4.11
	$\widehat{\mu}_6$	13.47	10.51	9.37	14.18	11.31	10.03	16.16	11.49	9.78	18.34	13.35	11.88

Table 4. Estimation of conditional standard deviation in continuous design; 10,000 replications;